Data Min Knowl Disc
https://doi.org/10.1007/s10618-018-0558-x

SICA: subjectively interesting component analysis

Bo Kang[1] · **Jefrey Lijffijt**[1] · **Raúl Santos-Rodríguez**[2] · **Tijl De Bie**[1]

Received: 2 January 2017 / Accepted: 23 February 2018

Abstract The information in high-dimensional datasets is often too complex for human users to perceive directly. Hence, it may be helpful to use dimensionality reduction methods to construct lower dimensional representations that can be visualized. The natural question that arises is *how do we construct a most informative low dimensional representation?* We study this question from an information-theoretic perspective and introduce a new method for linear dimensionality reduction. The obtained model that quantifies the informativeness also allows us to flexibly account for prior knowledge a user may have about the data. This enables us to provide representations that are *subjectively interesting*. We title the method Subjectively Interesting Component Analysis (SICA) and expect it is mainly useful for iterative data mining. SICA

Responsible editor: Fei Wang.

This work has received Funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement No. 615517, from the FWO (Project Numbers G091017N, G0F9816N) from the European Union's Horizon 2020 research and innovation programme and the FWO under the Marie Skłodowska-Curie Grant Agreement No. 665501, and from the EPSRC (EP/M000060/1).

✉ Bo Kang
Bo.Kang@ugent.be

Jefrey Lijffijt
Jefrey.Lijffijt@ugent.be

Raúl Santos-Rodríguez
enrsr@bristol.ac.uk

Tijl De Bie
Tijl.DeBie@ugent.be

[1] Department of Electronics and Information Systems, IDLab, Ghent University, Ghent, Belgium

[2] Data Science Lab, University of Bristol, Bristol, UK

Published online: 08 March 2018

is based on a model of a user's belief state about the data. This belief state is used to search for surprising views. The initial state is chosen by the user (it may be empty up to the data format) and is updated automatically as the analysis progresses. We study several types of prior beliefs: if a user only knows the scale of the data, SICA yields the same cost function as Principal Component Analysis (PCA), while if a user expects the data to have outliers, we obtain a variant that we term t-PCA. Finally, scientifically more interesting variants are obtained when a user has more complicated beliefs, such as knowledge about similarities between data points. The experiments suggest that SICA enables users to find subjectively more interesting representations.

Keywords Exploratory data mining · Dimensionality reduction · Information theory · Subjective interestingness · FORSIED

1 Introduction

The amount of information in high dimensional data makes it impossible to interpret such data directly. However, the data can be analyzed in a controlled manner, by revealing particular perspectives of data (lower dimensional data representations), one at a time. This is often done by means of projecting the data from the original feature space into a lower-dimensional subspace. Hence, such lower dimensional representations of a dataset are also called *data projections*, which are computed by a dimensionality reduction (DR) method.

DR methods are widely used for a number of purposes. The most prominent are data compression, feature construction, regularization in prediction problems, and exploratory data analysis. The most widely known DR technique, Principal Component Analysis (PCA) (Peason 1901) is used for each of these purposes (Bishop 2006), since it is computationally efficient, and more importantly, because large variance is often associated with structure, while noise often has smaller variance.

Other DR methods include linear methods such as Multidimensional Scaling (Kruskal and Wish 1978), Independent Component Analysis (Hyvärinen et al. 2004) and Canonical Correlations Analysis (Hotelling 1936), and non-linear techniques such as ISOMAP (Tenenbaum et al. 2000), Locality Preserving Projections (He and Niyogi 2004), and Laplacian-regularized models (Weinberger et al. 2006). The aforementioned methods all have objective score functions whose optimization yields the lower-dimensional representation, and they do not involve human users directly. Hence, we argue that these methods may well be suitable for, e.g., compression or regularization, but not optimal for providing most insight.

In exploratory data analysis, data is often visualized along the dimensions given by a DR method. Humans are unmatched in spotting visual patterns but inefficient at crunching numbers. Hence, visualizing high dimensional data in human perceivable yet computer-generated 2D/3D space can efficiently help users to understand different perspectives of the data (Puolamaki et al. 2010). However, since different human operators have different prior knowledge and interests, they are unlikely to have equal interest in the same aspect of data. For instance, PCA might be applied to obtain an

impression about the spread of data. But for many users, the structure in the data with largest variance may not be relevant at all.

To address this issue, Projection Pursuit (PP) (Friedman and Tukey 1974) was proposed, which finds data projections according to a certain interestingness measure (IM), designed with specific goals. With the ability to choose between different IMs, PP balances the computational efficiency and its applicability. However, because there are many analysis tasks and users, very many IMs are required, and this has led to an explosion in the number of IMs. Hence, unlike DR used for a specific analysis task or a predictive model, it seems to be conceptually challenging to define a generic quality metric for DR in the tasks of exploratory data analysis. This is precisely the focus of this paper.

In this paper we present Subjectively Interesting Component Analaysis (SICA), a dimensionality reduction method that finds *subjectively interesting* data projections. That is, projections that are aimed to be interesting to a particular user. In order to do so, SICA relies on quantifying how interesting a data projection is to the user. This quantification is based on information theory and follows the principles of FORSIED (De Bie 2011). Here we discuss the central idea of FORSIED and more detail will follow in Sect. 2.

FORSIED is a data mining framework for quantifying *subjective interestingness of patterns*. The central idea is that a user's belief state about the dataset is modelled as a Maximum Entropy (MaxEnt) probability distribution over the space of possible datasets. This probability distribution is called the *background distribution* and is updated as the analysis progresses, based on user interaction and the patterns in the data provided to the user. One can quantify the probability that a given pattern is present in data that is randomly drawn from the background distribution. Clearly, the smaller this probability, the more surprising the pattern is, and the more information it conveys to the user. More specifically, in FORSIED, the self-information of the pattern, defined as minus the logarithm of that probability, is then proposed as a suitable measure of how informative it is given the belief state.

In this paper, we define a pattern syntax called *projection patterns* for data projections that is compatible with FORSIED. By following FORSIED's principles, we can quantify the probability of a projection given the user's belief state. The lower the probability, the more surprising and interesting the pattern is, since surprising information about the data is typically what is truly interesting (Hand et al. 2001). Because this surprisal is evaluated with respect to the belief state, SICA can evaluate the subjective interestingness of projection patterns with respect to a particular user.

Contributions We introduce SICA, a dimensionality reduction method that tracks a user's belief about the data and presents subjectively interesting data projections to the user. To achieve this,

- we define *projection patterns*, a pattern syntax for data projections (Sect. 2);
- we derive a measure that quantifies the *subjective interestingness* of projection patterns (Sect. 2);
- we propose a method that finds the most subjectively interesting projections in terms of an optimization problem (Sect. 2);
- we define three types of prior beliefs a user may have knowledge about (Sect. 3);

– we demonstrate that with different prior belief types, SICA is able to (approximately/exactly) find the subjectively most interesting patterns. In particular, for some prior belief types, the subjective interestingness can be efficiently optimized by solving an eigenvalue problem (Sect. 3);
– we present three case studies and investigate the practical advantages and drawbacks of our method, which show that it can be meaningful to account for available prior knowledge about the data (Sect. 4).

This paper is an integrated and extended version of papers by De Bie et al. (2016) and Kang et al. (2016).

2 Subjectively interesting component analysis

SICA allows one to find data projections that reveal unexpected variation in the data. In this section, we introduce the ingredients needed to achieve this. Namely, we (a) define an interestingness measure (IM) that quantifies the amount of information a projection conveys to a particular user, (b) following to the IM, find interesting data projections for the user. In Sect. 3, we then develop SICA further for various types of prior beliefs.

2.1 Notation

We use upper case bold face letters to denote matrices, lower case bold face letters for vectors, and normal lower case letters for scalars. We denote a d-dimensional real-valued dataset as $\hat{\mathbf{X}} \triangleq (\hat{\mathbf{x}}_1', \hat{\mathbf{x}}_2', \ldots, \hat{\mathbf{x}}_n')' \in \mathbb{R}^{n \times d}$, and the corresponding random variable as $\mathbf{X}$. We will refer to $\mathbb{R}^{n \times d}$, the space the data is known to belong to, as the *data space*. Dimensionality reduction methods search weight vectors $\mathbf{w} \in \mathbb{R}^d$ of unit norm (i.e. $\mathbf{w}'\mathbf{w} = 1$) onto which the data is projected by computing $\hat{\mathbf{X}}\mathbf{w}$. If k vectors are sought, they will be stored as columns of a matrix $\mathbf{W} \in \mathbb{R}^{d \times k}$. We will denote the projections of a data set $\hat{\mathbf{X}}$ onto the column vectors of $\mathbf{W}$ as $\hat{\mathbf{\Pi}}_{\mathbf{W}} \in \mathbb{R}^{n \times k}$, or formally: $\hat{\mathbf{\Pi}}_{\mathbf{W}} \triangleq \hat{\mathbf{X}}\mathbf{W}$, and analogously for the random variable counterpart $\mathbf{\Pi}_{\mathbf{W}} \triangleq \mathbf{X}\mathbf{W}$. We will write $\mathbf{I}$ to denote the identity matrix of appropriate dimensions, and $\mathbf{1}_{n \times d}$ (or $\mathbf{1}$ for short if the dimensions are clear from the context) to denote a n-by-d matrix with all elements $\mathbf{1}_{ij} = 1$. We define the matrix interval with lower bound $\mathbf{B}$ and upper bound $\mathbf{C}$ denoted by $\mathbf{A}_{n \times m} \in [\mathbf{B}_{n \times m}, \mathbf{C}_{n \times m}]$, which indicates $a_{i,j} \in [b_{i,j}, c_{i,j}]$ for every $i, j = \{1, 2, \ldots, n\} \times \{1, 2, \ldots, m\}$.

2.2 Subjective interestingness measure for projections

We now derive an IM for SICA following the framework for subjective interestingness measures (FORSIED) (De Bie 2011, 2013). FORSIED is a data mining framework that specifies on an abstract level how to model a user's belief state about a given dataset, and how to quantify the informativeness of patterns with respect to a particular user. It works as follows: in order to measure the subjective interestingness of projections, SICA needs to maintain a model of the user's belief state. In addition, SICA should be

able to describe data projections in a pattern syntax compatible with FORSIED. We discuss both these issues in turn below.

2.2.1 Modeling the user's belief state

We formalize a user's belief state as a probability distribution over the data space (De Bie 2011):

Definition 1 (*Background distribution*) The *background distribution* is a distribution over the data space $\mathbb{R}^{n\times d}$ that represents the user's belief state: the probability it assigns to any measurable subset of $\mathbb{R}^{n\times d}$ corresponds to the probability that the user would ascribe to the data $\hat{\mathbf{X}}$ belonging to that subset. The background distribution can be represented by a probability density function $p_{\mathbf{X}} : \mathbb{R}^{n\times d} \to \mathbb{R}^{+}$.

For brevity, and slightly abusively, we will often refer to the density function $p_{\mathbf{X}}$ as the background distribution.

Of course, the background distribution is typically not known to the data mining system. Thus, it has to be inferred from limited information provided by the user. De Bie (2013) proposed an intuitive while mathematically rigorous language a user can employ to express certain beliefs about the data. The language assumes that important characteristics of the data can be quantified by means of statistics $f : \mathbb{R}^{n\times d} \to \mathbb{R}$. Using such statistics, the user can express their beliefs by declaring which value they expect f to have when evaluated on the data. Mathematically, this then becomes a constraint on the background distribution $p_{\mathbf{X}}$.

Definition 2 (*Prior belief constraints*) When the user expresses a prior belief by declaring that they expect a specified statistic f to be equal to a specified value $\hat{m} \in \mathbb{R}$, they are declaring that their background distribution $p_{\mathbf{X}}$ satisfies the following prior belief constraint:

$$\mathbb{E}_{\mathbf{X}\sim p_{\mathbf{X}}}[f(\mathbf{X})] = \hat{m}. \tag{1}$$

Except in degenerate cases, such constraints will not uniquely determine $p_{\mathbf{X}}$, such that an additional criterion is required to decide which one to use. Amongst those satisfying the prior belief constraints, the distribution with the maximum entropy (MaxEnt) is an attractive choice, given its unbiasedness and robustness. Further, as the resulting distribution belongs to the *exponential family*, its inference is well understood and often computationally tractable.

Formally, a user's background distribution can thus be obtained by solving the following constrained entropy maximization problem:

$$\begin{aligned} \underset{p_{\mathbf{X}}(\mathbf{X})\geq 0}{\text{argmax}} \quad & -\int p_{\mathbf{X}}(\mathbf{X}) \log\left(p_{\mathbf{X}}(\mathbf{X})\right) d\mathbf{X} \\ \text{s.t.} \quad & \int p_{\mathbf{X}}(\mathbf{X}) f_i(\mathbf{X}) d\mathbf{X} = \hat{m}_i, \quad \forall i, \\ & \int p_{\mathbf{X}}(\mathbf{X}) d\mathbf{X} = 1. \end{aligned} \tag{2}$$

As we will show in Sect. 3, by solving optimization problem (2) with different types of statistics f_i, one can model a wide variety of prior beliefs, and hence obtain very different types of background distributions.

2.2.2 Projection patterns: a pattern syntax for data projections

In FORSIED,[1] a *pattern* is defined as any information that restricts the set of possible values the data may have. For example, if the user is shown a scatter plot of the projections in $\hat{\mathbf{\Pi}}_{\mathbf{W}}$, the user will from then on know that $\hat{\mathbf{X}}\mathbf{W}$ is equal to $\hat{\mathbf{\Pi}}_{\mathbf{W}}$ (up to the resolution of the plot), which clearly constrains the set of possible values of the data to a subset of $\mathbb{R}^{n\times d}$.

One could thus be tempted to define a *projection pattern* as a statement of the kind $\hat{\mathbf{X}}\mathbf{W} = \hat{\mathbf{\Pi}}_{\mathbf{W}}$. This would tell the user that the projections of the data $\hat{\mathbf{X}}$ onto the columns of $\mathbf{W}$ are found to be equal to the columns of $\hat{\mathbf{\Pi}}_{\mathbf{W}}$.

However, real-valued data projections are often conveyed visually to a user, and in any case with finite accuracy, e.g. by means of a scatter plot. Because human eyes as well as the visualization devices (e.g., monitor, projector, and paper) have finite resolution, the precise value of the projected data can only be determined up to a certain resolution-dependent uncertainty $2\Delta\mathbf{1} \in \mathbb{R}^{n\times k}$. With these considerations[2], we formally define the syntax of a projection pattern as follows:

Definition 3 (*Projection pattern*) Let $\mathbf{W} \in \mathbb{R}^{d\times k}$ be a projection matrix, and let $\hat{\mathbf{\Pi}}_{\mathbf{W}}$ be the value of the projections of the data $\hat{\mathbf{X}} \in \mathbb{R}^{n\times d}$ onto the columns of $\mathbf{W}$. Then a *projection pattern* is a statement of the form:

$$\hat{\mathbf{X}}\mathbf{W} \in \left[\hat{\mathbf{\Pi}}_{\mathbf{W}} - \Delta\mathbf{1}, \hat{\mathbf{\Pi}}_{\mathbf{W}} + \Delta\mathbf{1}\right]. \tag{3}$$

Thus, the projection pattern specifies, up to an accuracy of 2Δ, the value of the projections of the data onto the columns of the projection matrix $\mathbf{W}$.

2.2.3 Subjective interestingness of projections

Relying on the background distribution, we can now quantify the *subjective interestingness* of a projection pattern:

[1] As well as in the only other framework for interactive data mining, CORAND (Lijffijt et al. 2014). By a framework for interactive data mining we mean a generic method that can be used to design specific data mining methods that take into account results previously shown to the user or other prior knowledge about the data. Such a framework would specify certain aspects of the method while other aspects are left open and only a guideline is provided on how to fill in that part. E.g., FORSIED specifies to define the background model as a MaxEnt distribution and the objective to maximize is the Subjective Interestingness. CORAND mandates another objective score (to maximize the p value of the data) and the form of the background distribution is left open; it may be anything. As far as we know, there are no other works published with a similar spirit.

[2] To simplify our notation, we assume the resolution parameter being the same through all dimensions. It is indeed an interesting direction to further develop SICA for the resolution varying in different dimensions.

Definition 4 (*Subjective interestingness of projection pattern*) The *subjective interestingness* (SI) of a projection pattern is defined as the negative log probability of the pattern under the background distribution.[3] For a projection pattern with projection matrix $\mathbf{W}$ and observed projections $\hat{\mathbf{\Pi}}_\mathbf{W}$, it is equal to:

$$\mathrm{SI}(\mathbf{W}, \hat{\mathbf{\Pi}}_\mathbf{W}) = -\log\left(\Pr\left(\mathbf{XW} \in \left[\hat{\mathbf{\Pi}}_\mathbf{W} - \Delta\mathbf{1}, \hat{\mathbf{\Pi}}_\mathbf{W} + \Delta\mathbf{1}\right]\right)\right). \tag{4}$$

The probability of a pattern can be computed by integrating the background distribution over all $\mathbf{X}$ for which $\mathbf{XW} \in \left[\hat{\mathbf{\Pi}}_\mathbf{W} - \Delta\mathbf{1}, \hat{\mathbf{\Pi}}_\mathbf{W} + \Delta\mathbf{1}\right]$:

$$\Pr\left(\mathbf{XW} \in \left[\hat{\mathbf{\Pi}}_\mathbf{W} - \Delta\mathbf{1}, \hat{\mathbf{\Pi}}_\mathbf{W} + \Delta\mathbf{1}\right]\right) = \int_{\mathbf{X}:\mathbf{XW}\in\left[\hat{\mathbf{\Pi}}_\mathbf{W}-\Delta\mathbf{1}, \hat{\mathbf{\Pi}}_\mathbf{W}+\Delta\mathbf{1}\right]} p_\mathbf{X}(\mathbf{X})d\mathbf{X}. \tag{5}$$

This can be expressed more conveniently in terms of the marginal density function $p_{\mathbf{\Pi}_\mathbf{X}}$ for the projection $\mathbf{\Pi}_\mathbf{W} \triangleq \mathbf{XW}$ of the data:

$$\Pr\left(\mathbf{XW} \in \left[\hat{\mathbf{\Pi}}_\mathbf{W} - \Delta\mathbf{1}, \hat{\mathbf{\Pi}}_\mathbf{W} + \Delta\mathbf{1}\right]\right) = \int_{\hat{\mathbf{\Pi}}_\mathbf{W}-\Delta\mathbf{1}}^{\hat{\mathbf{\Pi}}_\mathbf{W}+\Delta\mathbf{1}} p_{\mathbf{\Pi}_\mathbf{W}}(\mathbf{\Pi}_\mathbf{W})d\mathbf{\Pi}_\mathbf{W}. \tag{6}$$

For sufficiently small Δ, we approximate the integral in Eq. (6) as the value of the integrand in the middle of the integration domain times the integration domain's volume:[4]

$$\Pr\left(\mathbf{XW} \in \left[\hat{\mathbf{\Pi}}_\mathbf{W} - \Delta\mathbf{1}, \hat{\mathbf{\Pi}}_\mathbf{W} + \Delta\mathbf{1}\right]\right) \approx p_{\mathbf{\Pi}_\mathbf{W}}(\hat{\mathbf{\Pi}}_\mathbf{W})(2\Delta)^{nk}. \tag{7}$$

Then Definition 4 can be reformulated into:

$$\mathrm{SI}(\mathbf{W}, \hat{\mathbf{\Pi}}_\mathbf{W}) \approx -\log\left(p_{\mathbf{\Pi}_\mathbf{W}}(\hat{\mathbf{\Pi}}_\mathbf{W})\right) - nk\log(2\Delta). \tag{8}$$

Thus, to compute the interestingness of a projection pattern it is sufficient to know the marginal density function $p_{\mathbf{\Pi}_\mathbf{W}}$. We will compute this marginal density function in Sect. 3 for a number of background distributions.

2.3 Searching subjectively interesting projection patterns

Searching for subjectively interesting projection patterns amounts to finding a set of weight vectors $\mathbf{W} \in \mathbb{R}^{d\times k}$ that yield projections with the largest SI value. The

[3] In FORSIED, the subjective interestingness of a pattern is generally defined by a trade off between the information content (i.e., negative probability) of the pattern and the descriptional complexity (i.e., the amount of effort needed to assimilate the pattern). Here we assume all projections of the same dataset have the same descriptional complexity. As a result, the descriptional complexity can be ignored from the definition of SI.

[4] The tightness of this approximation for the cases in Sects. 3.1 and 3.2 will be investigated in detail in "Appendices A and B".

resulting weight vectors $\mathbf{W}$ linearly transform the original d features of the data $\hat{\mathbf{X}}$ into k features. Similar to the definition of the (principal) components in PCA, we refer to those k transformed features as the *subjectively interesting components* (SICs) of the data $\hat{\mathbf{X}}$.

The projection matrix $\mathbf{W}$ that corresponds to the SICs of data $\hat{\mathbf{X}}$ under background distribution $p_{\mathbf{X}}$ can thus be obtained by finding the $\mathbf{W}$ maximizing $\mathrm{SI}(\mathbf{W}, \hat{\mathbf{\Pi}}_{\mathbf{W}})$. As $\hat{\mathbf{X}}\mathbf{W}$ must represent projections, all weight vectors in the columns $\mathbf{W}$ must have unit norm. Additionally, to ensure non-redundancy of the different projections, we will require the weight vectors to be orthogonal, such that $\mathbf{W}'\mathbf{W} = \mathbf{I}$. Substituting $\hat{\mathbf{X}}\mathbf{W}$ for $\hat{\mathbf{\Pi}}_{\mathbf{W}}$ to make the dependencies on the data $\hat{\mathbf{X}}$ and the projection matrix $\mathbf{W}$ explicit, the optimization problem to be solved is thus:

$$\begin{aligned} \underset{\mathbf{W}\in\mathbb{R}^{d\times k}}{\text{argmax}} \quad & -\log\left(p_{\mathbf{\Pi}_{\mathbf{W}}}\left(\hat{\mathbf{X}}\mathbf{W}\right)\right), \\ \text{s.t.} \quad & \mathbf{W}'\mathbf{W} = \mathbf{I}. \end{aligned} \tag{9}$$

Note that this problem is independent of the resolution parameter Δ. In other words, as soon as Δ is small enough for Eq. 7 to hold to a sufficient approximation, its precise value is irrelevant to the problem.

It is this problem that we will be solving in Sect. 3 for a number of different types of background distributions.

3 SICA with different types of prior beliefs

In this section, we develop SICA further for three different types of prior beliefs. Each is discussed in a separate subsection. In Sect. 3.4, we discuss how SICA can in principle be used for other prior belief types as well, while also highlighting the difficulties in tackling other prior belief types that may limit the applicability of SICA in practice.

3.1 Scale of the data as prior belief

When the user only has a prior belief about the average variance of a dataset, SICA will aim to find projections with large variances. As we will show here, SICA with such prior is equivalent to PCA.

Prior belief With a given dataset, the user might have certain prior knowledge about the scale of a dataset. She might believe that the average scale of the data points, quantified by their squared norms, is some constant $\sigma^2 d$ and have no other knowledge. This can be formalized in a constraint of the form of Eq. (1):

$$\mathbb{E}_{\mathbf{X}\sim p_{\mathbf{X}}}\left[\frac{1}{n}\mathrm{Tr}\left(\mathbf{X}\mathbf{X}'\right)\right] = \sigma^2 d. \tag{10}$$

The corresponding statistic f of prior (10) is $f(\mathbf{X}) = \frac{1}{n}\mathrm{Tr}\left(\mathbf{X}\mathbf{X}'\right)$.

Background distribution Inserting (10) into (2), we obtain the following MaxEnt problem:

$$\begin{aligned} \underset{p_{\mathbf{X}}(\mathbf{X})\geq 0}{\operatorname{argmax}} \; & -\int p_{\mathbf{X}}(\mathbf{X}) \log\left(p_{\mathbf{X}}(\mathbf{X})\right) d\mathbf{X} \\ \text{s.t.} \; & \int p_{\mathbf{X}}(\mathbf{X}) \cdot \frac{1}{n} \operatorname{Tr}\left(\mathbf{X}\mathbf{X}'\right) d\mathbf{X} = \sigma^2 d, \\ & \int p_{\mathbf{X}}(\mathbf{X}) d\mathbf{X} = 1. \end{aligned} \tag{11}$$

The optimal background distribution $p_{\mathbf{X}}$ is a product distribution of identical multivariate Normal distributions with mean $\mathbf{0}$ and covariance matrix $\sigma^2\mathbf{I}$. This is summarized in the following theorem:

Theorem 1 *Given prior belief* (10), the MaxEnt background distribution is

$$p_{\mathbf{X}}(\mathbf{X}) = \prod_{i=1}^{n} p_{\mathbf{x}}(\mathbf{x}_i), \tag{12}$$

where $p_{\mathbf{x}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi\sigma^2)^d}} \exp\left(-\frac{\mathbf{x}'\mathbf{x}}{2\sigma^2}\right)$ is multivariate Normal density function with mean zero and covariance matrix $\sigma^2\mathbf{I}$.

Proof Through application of the Lagrange multiplier method, we find the Lagrangian for Problem (11):

$$\begin{aligned} \mathcal{L}\left(p_{\mathbf{X}}, \lambda, \mu\right) = & -\int p_{\mathbf{X}}\left(\mathbf{X}\right) \log\left(p_{\mathbf{X}}(\mathbf{X})\right) d\mathbf{X} - \lambda\left(\int p_{\mathbf{X}}(\mathbf{X}) \cdot \frac{1}{n} \operatorname{Tr}(\mathbf{X}\mathbf{X}') d\mathbf{X} - \sigma^2 d\right) \\ & + \mu\left(\int p_{\mathbf{X}}(\mathbf{X}) d\mathbf{X} - 1\right). \end{aligned} \tag{13}$$

Then, finding the function $p_{\mathbf{X}}$ that maximize this functional amounts to solving a *Euler–Lagrange equation* with Lagrangian $\mathcal{L}$ in form (13):

$$\frac{\partial \mathcal{L}}{\partial p_{\mathbf{X}}} - \frac{d}{d\mathbf{X}} \frac{\partial \mathcal{L}}{\partial p'_{\mathbf{X}}} = \frac{\partial \mathcal{L}}{\partial p_{\mathbf{X}}} + 0 = 0. \tag{14}$$

Hence, we compute the functional derivative of the Lagrangian with respect to $p_{\mathbf{X}}$ at $\mathbf{X}$:

$$\frac{\partial}{\partial p_{\mathbf{X}}(\mathbf{X})} \mathcal{L} = -1 - \log\left(p_{\mathbf{X}}(\mathbf{X})\right) + \frac{\lambda}{n} \operatorname{Tr}(\mathbf{X}\mathbf{X}') + \mu. \tag{15}$$

By equating the partial derivative to zero, we obtain an expression of $p_{\mathbf{X}}$ parametrized by λ and μ:

$$\begin{aligned} p_{\mathbf{X}}(\mathbf{X}) &= \exp\left(\mu - 1 + \frac{\lambda}{n}\mathrm{Tr}(\mathbf{X}\mathbf{X}')\right) \\ &= \exp(\mu - 1)\exp\left(\frac{\lambda}{n}\sum_{i=1}^{n}\mathbf{x}_i'\mathbf{x}_i\right) \\ &= \prod_{i=1}^{n}\frac{1}{Z}\exp\left(\frac{\lambda}{n}\mathbf{x}_i'\mathbf{x}_i\right), \end{aligned} \tag{16}$$

where $Z = \exp^{-1}\left(\frac{\mu-1}{n}\right)$. In order to find optimal solutions for λ and μ, observe that $p_{\mathbf{X}}(\mathbf{X})$ in Eq. (16) is the product of n identical multivariate Normal distributions, one for each data point $\mathbf{x}_i$, with zero mean and $-\frac{n}{2\lambda}\mathbf{I}$ as covariance matrix. As the expected two-norm squared of a multivariate Normal random vector with zero mean is equal to the trace of its covariance matrix, and as the expectation of the average two-norm squared of the identically distributed data points is constrained to be $\sigma^2 d$, this means that $\sigma^2 d = -\frac{dn}{2\lambda}$, such that $\lambda = -\frac{n}{2\sigma^2}$.

Therefore, the MaxEnt background distribution is an independent multivariate normal distribution, where each independent random variable has zero mean, and covariance matrix $\sigma^2\mathbf{I}$, i.e., $\mathcal{N}\left(\mathbf{0}, \sigma^2\mathbf{I}\right)$. □

Subjectively interesting patterns Now we can search for subjectively interesting patterns by solving problem (9). This requires to first compute distribution $p_{\mathbf{\Pi}_\mathbf{W}}$ as the marginal of the background distribution (16).

Given a normal random vector $\mathbf{x} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2\mathbf{I}\right)$, a projection onto weight vectors $\mathbf{W}$ with $\mathbf{W}'\mathbf{W} = \mathbf{I}$ is also normal: $\mathbf{x}'\mathbf{W} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2\mathbf{I}\right)$. Thus, the marginal density function distribution $p_{\mathbf{\Pi}_\mathbf{W}}$ for the projection $\mathbf{\Pi}_\mathbf{W} = \mathbf{X}\mathbf{W}$ is given by:

$$\begin{aligned} p_{\mathbf{\Pi}_\mathbf{W}}(\mathbf{X}\mathbf{W}) &= \prod_{i=1}^{n}\prod_{j=1}^{k}\frac{1}{\sqrt{(2\pi\sigma^2)}}\exp\left(-\frac{(\mathbf{w}_j'\mathbf{x}_i)^2}{2\sigma^2}\right) \\ &= \prod_{i=1}^{n}\frac{1}{\sqrt{(2\pi\sigma^2)^k}}\exp\left(-\frac{\mathbf{x}_i'\mathbf{W}\mathbf{W}'\mathbf{x}_i}{2\sigma^2}\right) \\ &= \frac{1}{\sqrt{(2\pi\sigma^2)^{nk}}}\exp\left(-\frac{1}{2\sigma^2}\mathrm{Tr}\left[\mathbf{W}'\mathbf{X}'\mathbf{X}\mathbf{W}\right]\right). \end{aligned} \tag{17}$$

Given density function (17), we can now use (9) to find projection patterns ($\mathbf{X}\mathbf{W} \in [\hat{\mathbf{\Pi}}_\mathbf{W} - \Delta\mathbf{1}, \hat{\mathbf{\Pi}}_\mathbf{W} + \Delta\mathbf{1}]$) that are subjectively interesting. This is only true if the approximation (7) is good enough. In "Appendix A", we show this is indeed the case. Thus, substituting the marginal distribution (17) into the objective function of problem (9) gives:

$$-\log\left(p_{\mathbf{\Pi}_\mathbf{W}}(\hat{\mathbf{\Pi}}_\mathbf{W})\right) = \frac{nk}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\mathrm{Tr}\left[\mathbf{W}'\hat{\mathbf{X}}'\hat{\mathbf{X}}\mathbf{W}\right]. \tag{18}$$

Ignoring the first constant term and constant factor $\frac{1}{2\sigma^2}$, the optimization problem (9) is equivalent to:

$$\begin{aligned} \max_{\mathbf{W}\in\mathbb{R}^{d\times k}} \ & \mathrm{Tr}\left[\mathbf{W}'\hat{\mathbf{X}}'\hat{\mathbf{X}}\mathbf{W}\right] \\ \text{s.t. } & \mathbf{W}'\mathbf{W} = \mathbf{I}. \end{aligned} \tag{19}$$

This is equivalent (up to rotation) to the problem of finding the k dominant principal component of $\mathbf{X}$ in classical PCA.[5]

3.2 t-PCA: magnitude of spread as prior belief

In contrast to believing the data has a certain scale, a user might expect that the data has certain magnitude of spread. In this subsection, we show that with such prior expectation, SICA yields an alternative result, that turns out to be more robust against outliers.

Prior belief Denote γ to be the parameter that expresses the user's belief about the magnitude of spread of the data. The user's expectation about the magnitude of spread to be some value a is then defined by:

$$\mathbb{E}_{\mathbf{X}\sim p_\mathbf{X}}\left[\frac{1}{n}\sum_{i=1}^{n}\log\left(1+\frac{1}{\gamma}\mathbf{x}_i'\mathbf{x}_i\right)\right] = a. \tag{20}$$

If the user is expecting outliers in the data, she may specify γ to be small. This will up-weight the outliers (who have relatively large 2-norms) such that they contribute more to the expectation. In contrast, by setting a larger γ, the expectation is focused more on the bulk of the points.

Background distribution with the prior belief (20) we need to solve the following optimization problem to obtain the MaxEnt background distribution:

[5] In this paper, by performing PCA, we mean the data $\mathbf{X}$ is first centered ($\mathbf{X}_c = \mathbf{X} - \frac{1}{n}\mathbf{1}_{n\times 1}\mathbf{1}_{n\times 1}'\mathbf{X}$), then the eigenvectors of matrix $\mathbf{X}'\mathbf{X}$ are computed and sorted in descending order according to the absolute value of the eigenvalues. After sorting, the eigenvectors of $\mathbf{X}'\mathbf{X}$ with largest absolute eigenvalues correspond to the top principal components.

$$\begin{aligned} \underset{p_{\mathbf{X}}(\mathbf{X})\geq 0}{\text{argmax}} \; & -\int p_{\mathbf{X}}(\mathbf{X}) \log\left(p_{\mathbf{X}}(\mathbf{X})\right) d\mathbf{X} \\ \text{s.t.} \; & \int p_{\mathbf{X}}(\mathbf{X}) \cdot \frac{1}{n} \sum_{i=1}^{n} \log\left(1+\frac{1}{\gamma}\mathbf{x}_i'\mathbf{x}_i\right) d\mathbf{X} = a, \\ & \int p_{\mathbf{X}}(\mathbf{X}) d\mathbf{X} = 1. \end{aligned} \tag{21}$$

Relying on the result by Zografos (1999), we find that the optimal solution is a product of independent multivariate standard t-distributions. Here, we denote a digamma function as φ, and introduce the function $\kappa(\nu) = \varphi(\frac{\nu+d}{2}) - \varphi(\frac{\nu}{2})$, where d is the dimension of data $\hat{\mathbf{X}}$. In the sequel, the value $\nu = \kappa^{-1}(a)$ will be used, i.e., ν depends on the expected magnitude of spread a. The background distribution with prior belief (20) is then defined as:

Theorem 2 *Given prior belief* (20)*, the MaxEnt background distribution is*

$$p_{\mathbf{X}}(\mathbf{X}) = \prod_{i=1}^{n} p(\mathbf{x}_i) \tag{22}$$

where $p(\mathbf{x})$ *is the density function of a multivariate standard t-distribution with form:*

$$p(\mathbf{x}) = \frac{\Gamma\left(\frac{\nu+d}{2}\right)}{(\pi\rho)^{d/2}\Gamma\left(\frac{\nu}{2}\right)} \cdot \frac{1}{\left(1+\frac{1}{\rho}\mathbf{x}'\mathbf{x}\right)^{\frac{\nu+d}{2}}}. \tag{23}$$

where $\rho = \gamma\nu$*, the correlation matrix is a* d*-by-*d *identity matrix* $\mathbf{I}$*.*

Proof We restate the Theorem 2.1 and the derivation of Eq. 2.12 from the paper by Zografos (1999). From this the proof immediately follows.

Theorem 2.1 in (Zografos 1999) sates that for MaxEnt problem:

$$\begin{aligned} \underset{p_{\mathbf{x}}(\mathbf{x})\geq 0}{\text{argmax}} \; & -\int p_{\mathbf{x}}(\mathbf{x}) \log\left(p_{\mathbf{x}}(\mathbf{x})\right) d\mathbf{x} \\ \text{s.t.} \; & \int p_{\mathbf{x}}(\mathbf{x}) \log\left(1+(\mathbf{x}-\mu)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mu)\right) d\mathbf{x} = \varphi(m) - \varphi\left(m-\frac{d}{2}\right) \\ & \int p_{\mathbf{x}}(\mathbf{x}) d\mathbf{x} = 1. \end{aligned}$$

where $\mathbf{x} \in \mathbb{R}^d$,$m > (d+2)/2$, $\mathbb{E}(\mathbf{x}) = \mu$, $\text{Cov}(\mathbf{x}) = 1/(2m-d-2)\boldsymbol{\Sigma}$. The solution of this problem is a special case of *Pearson's Type VII* multivariate distribution with density:

$$p(\mathbf{x}) = \frac{\Gamma(m)}{\pi^{d/2}\Gamma(m-d/2)} |\boldsymbol{\Sigma}|^{-1/2} [1+(\mathbf{x}-\mu)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mu)]^{-m}.$$

That is, the multivariate t-distribution with ν degrees of freedom, $\nu = \kappa^{-1}(a) > 0$, $\mu = \mathbf{0}$, and $\mathbf{\Sigma} = \gamma\mathbf{I}$ can be obtained from Pearson's Type VII distribution using transformation $\mathbf{z} = \sqrt{\nu}\mathbf{x} + (1 - \sqrt{\nu})\mu$ and taking $m = (\nu + d)/2$:

$$\begin{aligned} p(\mathbf{z}) &= \frac{\Gamma((\nu+d)/2)}{(\pi\nu\gamma)^{d/2}\Gamma(\nu/2)}\left[1 + \frac{1}{\nu\gamma}\mathbf{z}'\mathbf{z}\right]^{-(\nu+d)/2} \\ &= \frac{\Gamma((\nu+d)/2)}{(\pi\rho)^{d/2}\Gamma(\nu/2)}\left[1 + \frac{1}{\rho}\mathbf{z}'\mathbf{z}\right]^{-(\nu+d)/2} \end{aligned}$$

By setting $\rho = \gamma\nu$, only one parameter needs to be tuned. □

Remark 1 Note that for $\rho, \nu \to \infty$, $\frac{\rho}{\nu} \to \sigma^2$ this density function tends to the multivariate Normal density function with mean $\mathbf{0}$ and covariance $\sigma^2\mathbf{I}$. For $\rho = \nu = 1$ it is a multivariate standard Cauchy distribution, which is so heavy-tailed that its mean is undefined and its second moment is infinitely large. Thus, this type of prior belief can model the expectation of outliers to varying degrees.

Given the reliance on a multivariate t-distribution as the background distribution, we will refer to this model as *t-PCA*.

Subjectively interesting patterns According to Kotz and Nadarajah (2004), the marginals of a t-distribution with given correlation matrix are again a t-distribution with the same number of degrees of freedom. Each marginal is obtained by selecting the relevant part of the correlation matrix. This means that the marginal density function for projection $\mathbf{\Pi}_\mathbf{W} = \mathbf{XW}$ onto k weight vectors $\mathbf{W}$ with $\mathbf{W}'\mathbf{W} = \mathbf{I}$ is

$$p_{\mathbf{\Pi}_\mathbf{W}}(\mathbf{\Pi}_\mathbf{W}) = \prod_{i=1}^{n} \frac{\Gamma\left(\frac{\nu+k}{2}\right)}{(\pi\rho)^{k/2}\Gamma\left(\frac{\nu}{2}\right)} \cdot \frac{1}{\left(1 + \frac{1}{\rho}\mathbf{x}_i'\mathbf{W}\mathbf{W}'\mathbf{x}_i\right)^{\frac{\nu+k}{2}}}. \tag{24}$$

Given density function (24), we can now use (9) to find projection patterns ($\mathbf{XW} \in [\hat{\mathbf{\Pi}}_\mathbf{W} - \Delta\mathbf{1}, \hat{\mathbf{\Pi}}_\mathbf{W} + \Delta\mathbf{1}]$) that are subjectively interesting. This is only true if the approximation (7) is good enough. In "Appendix B", we show this is indeed the case. Thus, substituting the marginal distribution (24) into the objective function of problem (9) by gives:

$$-\log\left(p_{\mathbf{\Pi}_\mathbf{W}}(\hat{\mathbf{\Pi}}_\mathbf{W})\right) = \frac{\nu+k}{2}\sum_{i=1}^{n}\log\left(1 + \frac{1}{\rho}\hat{\mathbf{x}}_i'\mathbf{W}\mathbf{W}'\hat{\mathbf{x}}_i\right) + \text{a constant}. \tag{25}$$

Ignoring some constant factors and terms, searching for the subjectively most interesting pattern is thus equivalent to solve:

$$\begin{aligned} &\max_{\mathbf{W}\in\mathbb{R}^{d\times k}} \sum_{i=1}^{n}\log(\rho + \hat{\mathbf{x}}_i'\mathbf{W}\mathbf{W}'\hat{\mathbf{x}}_i) \\ &\text{s.t. } \mathbf{W}'\mathbf{W} = \mathbf{I}. \end{aligned} \tag{26}$$

Remark 2 By varying ρ, SICA interpolates between maximizing the arithmetic mean, like PCA does, and maximizing the geometric mean of the square of the data projections, which is more robust against outliers. To be precise, for $\rho = 0$ the objective function (26) is monotonically related to the geometric mean of the squared norm of data projections $\|\hat{\mathbf{x}}_i'\mathbf{W}\|^2$:

$$\exp\left[\frac{1}{n}\sum_{i=1}^{n}\log(\|\hat{\mathbf{x}}_i'\mathbf{W}\|^2)\right] = \left(\prod_{i=1}^{n}(\|\hat{\mathbf{x}}_i'\mathbf{W}\|^2)\right)^{\frac{1}{n}}. \tag{27}$$

On the other hand, for $\rho \to \infty$, the objective function (26) is monotonically related to arithmetic mean,

$$\begin{aligned}
&\lim_{\rho\to\infty}\frac{\rho}{n}\sum_{i=1}^{n}\log\left(\rho + \|\hat{\mathbf{x}}_i'\mathbf{W}\|^2\right) - \rho\log(\rho)\\
&= \lim_{\rho\to\infty}\frac{1}{n}\sum_{i=1}^{n}\rho\log\left(1 + \frac{\|\hat{\mathbf{x}}_i'\mathbf{W}\|^2}{\rho}\right) + \rho\log(\rho) - \rho\log(\rho)\\
&= \lim_{\rho\to\infty}\frac{1}{n}\sum_{i=1}^{n}\log\left[\left(1 + \frac{\|\hat{\mathbf{x}}_i'\mathbf{W}\|^2}{\rho}\right)^{\rho}\right]\\
&= \frac{1}{n}\sum_{i=1}^{n}\|\hat{\mathbf{x}}_i'\mathbf{W}\|^2,
\end{aligned}$$

That is, for sufficiently large ρ the objective function is equivalent to the arithmetic mean, up to factor ρ and additive constant $-\rho\log(\rho)$.

To get some insight into the computational complexity of problem (26), let us consider the one dimensional case where we search for weight vector $\mathbf{w} \in \mathbb{R}^{d\times 1}$. Clearly, the larger $\mathbf{w}'\mathbf{w}$, the larger the objective, so the constraint can be relaxed to $\mathbf{w}'\mathbf{w} \le 1$. Hence the feasible set of $\mathbf{w}$ is convex. Denote $s_i = \text{sign}\left(\hat{\mathbf{x}}_i\mathbf{w}\right)$ as the sign of the scale value $\hat{\mathbf{x}}_i\mathbf{w}$. For $\rho = 0$, the objective can be re-written as $\sum_{i=1}^{n}\log\left((\hat{\mathbf{x}}_i'\mathbf{w})^2\right) = \sum_{i=1}^{n}\log\det\begin{pmatrix} s_i\hat{\mathbf{x}}_i'\mathbf{w} & 0 \\ 0 & s_i\hat{\mathbf{x}}_i'\mathbf{w}\end{pmatrix}$, which is a sum of log determinant functions of the parameters $\mathbf{w}$. Hence the objective function is concave. Based on this observation, a possible solution strategy is to enumerate all possible sign vector $\mathbf{s} = \text{sign}(\hat{\mathbf{X}}\mathbf{w}_i)$, and first find an optimal $\mathbf{w}$ for each of those convex problems. The global optimal solution can then be found over all enumerations. Although this is not a proof of the complexity of the problem, and the existence of an efficient algorithm cannot be ruled out, it shows that at least a naive algorithm needs an exponential time in $\mathcal{O}((n-1)^{d-1})$.

We solve the problem (26) by approximation. Observe that the orthonormality constraint posed on $\mathbf{W}$ leads to problem (26) being a Stiefel manifold (Onishchik 2011) optimization problem. This can be addressed fairly efficiently with a standard tool box. We use the Manopt toolbox (Boumal et al. 2014) to obtain an approximate solution.

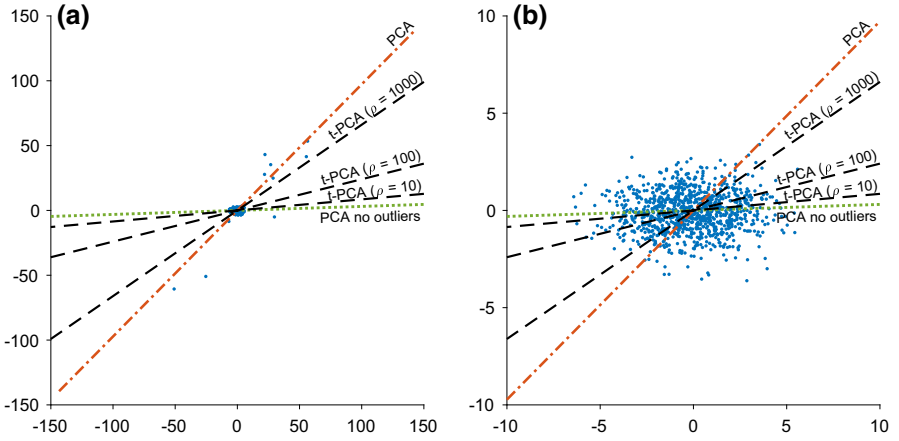

Fig. 1 Synthetic data (Sect. 3.2) visualized with weight vectors of PCA (red dash-dotted line), SICA (black dashed lines, $\rho = 10, 100, 1000$), and PCA fitted excluding the outliers (green dotted line). **a** Data visualized including outliers. **b** Data visualized excluding outliers

Remark 3 For the parameter ρ in constraint (20), a user can set it freely according to her prior belief. Namely, if the user feels confident about the average squared norm of the data points, a large ρ should be used, but if the user feels confident only about the *order of magnitude* of the norms of the data points, a small ρ should be used. The next example illustrates the effect of different choices for ρ.

Example As an illustrative example, we compare PCA and SICA on synthetic data. We generated a dataset consisting of two populations with different covariance structures: 1000 data points sampled from $\mathcal{N}\left(\mathbf{0}, \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix}\right)$, and 10 'outliers' from $\mathcal{N}\left(\mathbf{0}, \begin{pmatrix} 16 & 12 \\ 12 & 13 \end{pmatrix}\right)$, i.e., $\hat{\mathbf{X}} \in \mathbb{R}^{1010\times 2}$. After sampling, the data is centered. Figure 1a shows the first components resulting from PCA, SICA, and PCA had there been no outliers. The PCA result is determined primarily by the outliers. The right plot (Fig. 1b) shows the components on top of a scatter plot without the 10 outliers, illustrating that SICA is hardly affected by outliers. That is, the lower ρ the more the user's belief allows for the existence of outliers, hence SICA shows the projection with fewer outliers as additional information. By varying the ρ parameter ($\rho = 10, 100, 1000$), the resulted projection interpolates between PCA and PCA on data with outliers removed.

3.3 Pairwise data point similarities as prior beliefs

In SICA, users may specify not only global characteristics of the data, such as the expected magnitude of spread, but they can also express expectations about local characteristics, such as similarities between data points.

Prior belief Assume the user believes that a data point is similar to another point or group of points. She may then want to discover other structure within the data, in

addition to the known similarities. Generally speaking, the most interesting/surprising information would be a pattern that *contrasts* with the known similarities. For example, consider a user interested in social network analysis, and more specifically, interested in finding social groups that share certain properties. Suppose the user has already studied the network structure to some degree, and now it would be more interesting for her to learn about other properties shared by different social groups; other as in properties not aligned with the network structure.

SICA allows the user to encode their beliefs as follows. The data points are represented as nodes in a graph $G = (\mathbf{X}, E)$, and the user can connect all pairs of points that she expects to be similar with an edge. In this way, the user's prior belief regarding similarities among data points can be measured as the average pairwise Euclidean distance of connected nodes in graph G:

$$\mathbb{E}_{\mathbf{X}\sim p_{\mathbf{X}}}\left[\frac{1}{|E|}\sum_{(i,j)\in E}||\mathbf{x}_i - \mathbf{x}_j||^2\right] = b, \tag{28}$$

where b is some constant. Constraint (28) on its own still has ambiguity, as a small b can be due to a belief that connected data points in G are close together, but also due to a belief that the scale of the data is simply small. Thus, to forestall the second interpretation, another constraint needs to be imposed which fixes the expected scale of the data:

$$\mathbb{E}_{\mathbf{X}\sim p_{\mathbf{X}}}\left[\frac{1}{n}\sum_{i=1}^{n}||\mathbf{x}_i||^2\right] = c. \tag{29}$$

Background distribution To obtain the background distribution, the following MaxEnt problem needs to be solved:

$$\begin{aligned}
\underset{p_{\mathbf{X}}(\mathbf{X})\geq 0}{\text{argmax}} \; & -\int p_{\mathbf{X}}(\mathbf{X})\log\left(p_{\mathbf{X}}(\mathbf{X})\right)d\mathbf{X} \\
\text{s.t.} \; & \int p_{\mathbf{X}}(\mathbf{X})\cdot\frac{1}{|E|}\sum_{(i,j)\in E}||\mathbf{x}_i - \mathbf{x}_j||^2 d\mathbf{X} = b, \\
& \int p_{\mathbf{X}}(\mathbf{X})\cdot\frac{1}{n}\sum_{i=1}^{n}||\mathbf{x}_i||^2 d\mathbf{X} = c, \\
& \int p_{\mathbf{X}}(\mathbf{X})d\mathbf{X} = 1.
\end{aligned} \tag{30}$$

Denote $\mathbf{I}$ as identity matrix and $\mathbf{L}$ as the Laplacian of the graph G defined as $\mathbf{L} = \mathbf{D} - \mathbf{A}$, with $\mathbf{A}$ the adjacency matrix of graph and $\mathbf{D}$ the diagonal matrix with the degrees of nodes on its diagonal. We now show that the solution of problem (30) is a matrix normal distribution $\mathcal{MN}_{n\times d}(\mathbf{M}, \mathbf{\Sigma}, \mathbf{\Phi})$, specifically:

Theorem 3 *The optimal solution of problem* (30) *is given by a matrix normal distribution:*

$$\mathbf{X} \sim \mathcal{MN}_{n\times d}\left(\mathbf{0}, \left(2\left[\frac{\lambda_1}{|E|}\mathbf{L} + \frac{\lambda_2}{n}\mathbf{I}_n\right]\right)^{-1}, \mathbf{I}_d\right), \tag{31}$$

namely,

$$p_{\mathbf{X}}(\mathbf{X}) = \frac{1}{Z}\exp\left\{Tr\left(-\mathbf{X}'\left[\frac{\lambda_1}{|E|}\mathbf{L} + \frac{\lambda_2}{n}\mathbf{I}_n\right]\mathbf{X}\right)\right\}, \tag{32}$$

with partition function $Z = (2\pi)^{\frac{nd}{2}}\left|2\left[\frac{\lambda_1}{|E|}\mathbf{L} + \frac{\lambda_2}{n}\mathbf{I}_n\right]\right|^{\frac{d}{2}}$.

The proof, provided below, makes clear that the values of λ_1 and λ_2 depend on the values of b and c in the constraints, and can be found by solving a very simple convex optimization problem:

Proof The Lagrangian for (30) is:

$$\begin{aligned}\mathcal{L}(p_{\mathbf{X}}, \lambda, \mu) = &-\int p_{\mathbf{X}}(\mathbf{X})\log\left(p_{\mathbf{X}}(\mathbf{X})\right)d\mathbf{X} \\ &- \lambda_1\left(\int p_{\mathbf{X}}(\mathbf{X})\cdot\frac{1}{|E|}\sum_{(i,j)\in E}||\mathbf{x}_i - \mathbf{x}_j||^2 d\mathbf{X} - b\right) \\ &- \lambda_2\left(\int p_{\mathbf{X}}(\mathbf{X})\cdot\frac{1}{n}\sum_{i=1}^{n}||\mathbf{x}_i||^2 d\mathbf{X} - c\right) - \mu\left(\int p_{\mathbf{X}}(\mathbf{X})d\mathbf{X} - 1\right),\end{aligned} \tag{33}$$

whose partial derivative with respect to $p_{\mathbf{X}}$ at $\mathbf{X}$ reads:

$$\frac{\partial}{\partial p_{\mathbf{X}}(\mathbf{X})}\mathcal{L} = -1 - \log\left(p_{\mathbf{X}}(\mathbf{X})\right) - \frac{\lambda_1}{|E|}\sum_{(i,j)\in E}||\mathbf{x}_i - \mathbf{x}_j||^2 - \frac{\lambda_2}{n}\sum_{i=1}^{n}||\mathbf{x}_i||^2 - \mu. \tag{34}$$

Equating this partial derivative to zero yields:

$$\begin{aligned}p_{\mathbf{X}}(\mathbf{X}) &= \exp(-1-\mu)\cdot\exp\left\{-\frac{\lambda_1}{|E|}\sum_{(i,j)\in E}||\mathbf{x}_i - \mathbf{x}_j||^2 - \frac{\lambda_2}{n}\sum_{i=1}^{n}||\mathbf{x}_i||^2\right\} \\ &= \frac{1}{Z}\exp\left\{\mathrm{Tr}\left(-\mathbf{X}'\left[\frac{\lambda_1}{|E|}\mathbf{L} + \frac{\lambda_2}{n}\mathbf{I}_n\right]\mathbf{X}\right)\right\}.\end{aligned} \tag{35}$$

Observe that (35) is a matrix normal distribution (Gupta and Nagar 1999) with partition function Z and parameters $\mathbf{M} = 0$, $\mathbf{\Sigma} = \left(2\left[\frac{\lambda_1}{|E|}\mathbf{L} + \frac{\lambda_2}{n}\mathbf{I}_n\right]\right)^{-1}$, and $\mathbf{\Phi} = \mathbf{I}_d$. Hence, the matrix-valued random variable $\mathbf{X} \in \mathbb{R}^{n\times d}$ belongs to:

$$\mathbf{X} \sim \mathcal{MN}_{n\times d}\left(\mathbf{0}, \left(2\left[\frac{\lambda_1}{|E|}\mathbf{L} + \frac{\lambda_2}{n}\mathbf{I}_n\right]\right)^{-1}, \mathbf{I}_d\right), \tag{36}$$

with the partition function

$$Z = (2\pi)^{\frac{nd}{2}}\left|2\left[\frac{\lambda_1}{|E|}\mathbf{L} + \frac{\lambda_2}{n}\mathbf{I}_n\right]\right|^{\frac{d}{2}}. \tag{37}$$

□

Remark 4 To compute the multipliers λ_1 and λ_2, we substitute the distribution (35) back into the Lagrangian (33) and solve λ_1 and λ_2 that minimizes the following Lagrange dual function using, e.g., gradient based methods:

$$\mathcal{L}(\lambda) = \frac{d}{2}\log((2\pi)^n|\mathbf{\Sigma}|) + \lambda_1 b + \lambda_2 c,$$

where $\mathbf{\Sigma} = \left(\frac{2\lambda_1}{|E|}\mathbf{L} + \frac{2\lambda_2}{n}\mathbf{I}_n\right)^{-1}$. Since $\mathbf{L}$ is a real symmetric matrix, we can simultaneously diagonalize $\mathbf{L}$ and $\mathbf{I}_n$. Denote the eigenvalues of the matrix $\mathbf{L}$ to be $\sigma_1, \sigma_2, \ldots, \sigma_n$. Then the determinant of the covariance matrix reads:

$$|\mathbf{\Sigma}| = \prod_{i=1}^{n}\left(\frac{2\lambda_1\sigma_i}{|E|} + \frac{2\lambda_2}{n}\right)^{-1}.$$

Thus the Lagrange dual function can be further simplified as:

$$\mathcal{L}(\lambda) = -\frac{d}{2}\sum_{i=1}^{n}\log\left(\frac{2\lambda_1\sigma_i}{|E|} + \frac{2\lambda_2}{n}\right) + \frac{nd}{2}\log(2\pi) + \lambda_1 b + \lambda_2 c.$$

Hence, computing the multipliers requires to first compute the eigenvalues of $\mathbf{L}$ ($\mathcal{O}(n^3)$), then the evaluation of each gradient step has complexity $\mathcal{O}(n)$.

Remark 5 In order to determine suitable values for b and c in the prior belief constraints, SICA may assume that the user already has a good understanding of the point-wise similarity (Eq. 28) and scale (Eq. 29) of the data points (or, that the user is not interested in these). Given this assumption, b and c can simply be set equal to the empirical value of these statistics as measured in the data. If the user wishes, she could of course specify values herself that differ from these. More realistically though, she may be able to specify a *range* of values for the point-wise similarity and scale. The background distribution should then be found as the MaxEnt distribution subject to two *box constraints*, i.e., four inequality constraints: a lower and a upper bound for pairwise similarity as well as for the scale measure. Theorem 3 still applies unaltered though: while the four inequality constraints lead to four Lagrange multipliers, only two may be non-zero at the optimum (one for each box constraint), as for each box constraint only either the upper or the lower bound constraint can be tight.

Subjectively interesting patterns As the projection $\mathbf{\Pi_W}$ is a linear transformation of matrix random variable $\mathbf{X}$, and $\mathbf{W}$ is of rank $k \leq n$ (full column rank), then $\mathbf{\Pi_W} \sim \mathcal{MN}_{n\times k}\left(\mathbf{0}, \left(2\left[\frac{\lambda_1}{|E|}\mathbf{L} + \frac{\lambda_2}{n}\mathbf{I}_n\right]\right)^{-1}, \mathbf{I}_k\right)$ (Gupta and Nagar 1999). So the marginal $p_{\mathbf{\Pi_W}}$ of background distribution (31) reads:

$$p_{\mathbf{\Pi_W}}(\mathbf{\Pi_W}) = \frac{1}{Z}\exp\left\{\mathrm{Tr}\left(-\mathbf{\Pi}'_{\mathbf{W}}\left[\frac{\lambda_1}{|E|}\mathbf{L} + \frac{\lambda_2}{n}\mathbf{I}_n\right]\mathbf{\Pi_W}\right)\right\}. \tag{38}$$

Substituting the marginal distribution (38) into the objective function of problem (9), and $\hat{\mathbf{X}}\mathbf{W}$ for $\hat{\mathbf{\Pi}}_{\mathbf{W}}$, yields:

$$-\log\left(p_{\mathbf{\Pi_W}}\left(\hat{\mathbf{X}}\mathbf{W}\right)\right) = \mathrm{Tr}\left(\mathbf{W}'\hat{\mathbf{X}}'\left[\frac{\lambda_1}{|E|}\mathbf{L} + \frac{\lambda_2}{n}\mathbf{I}_n\right]\hat{\mathbf{X}}\mathbf{W}\right) + \log(Z). \tag{39}$$

Since the second term of (39) is constant, it can be safely left out. Thus the optimization problem (9) is equivalent to:

$$\begin{aligned} \max_{\mathbf{W}\in\mathbb{R}^{d\times k}} \ & \mathrm{Tr}\left(\mathbf{W}'\hat{\mathbf{X}}'\left[\frac{\lambda_1}{|E|}\mathbf{L} + \frac{\lambda_2}{n}\mathbf{I}_n\right]\hat{\mathbf{X}}\mathbf{W}\right) \\ \text{s.t. } & \mathbf{W}'\mathbf{W} = \mathbf{I}. \end{aligned} \tag{40}$$

The solution to this problem consists of a matrix $\mathbf{W} \in \mathbb{R}^{d\times k}$ whose k column vectors are the eigenvectors that corresponding to the top-k eigenvalues of the matrix $\hat{\mathbf{X}}'\left[\frac{\lambda_1}{|E|}\mathbf{L} + \frac{\lambda_2}{n}\mathbf{I}_n\right]\hat{\mathbf{X}} \in \mathbb{R}^{d\times d}$ (Kokiopoulou et al. 2011).

The computational complexity of finding an optimal projection $\mathbf{W}$ consists of two parts: (1) solving a convex optimization problem to obtain the background distribution. This can be achieved by applying, e.g., a steepest descent method, which uses at most $\mathcal{O}(\varepsilon^{-2})$ steps (until the norm of the gradient is $\leq \varepsilon$) (Nesterov 2013). For each step, the complexity is $\mathcal{O}(n)$ with n being the size of data. (2) Given the background distribution, we find an optimal projection, the complexity of which is dominated by eigenvalue decomposition ($\mathcal{O}(n^3)$). Hence, the overall complexity of SICA with graph prior is $\mathcal{O}(\frac{n}{\varepsilon^2} + n^3)$.

Example We synthesized a dataset with 100 users, where each user is described by 10 attributes, i.e., $\hat{\mathbf{X}} \in \mathbb{R}^{100\times 10}$. The first attribute is generated from a bimodal Gaussian distribution such that the first attribute clearly separates the users into two groups. We assume that people within each community are fully connected. To have a more interesting simulation, we also insert a few connections between the communities. The second attribute value is uniformly drawn from $\{-1, +1\}$ which could resemble, e.g., people's sentiment towards a certain topic. The remaining eight attributes are standard Gaussian noise. After sampling, we centered the data.

We assume the user has studied the observed connection between all data points. Hence, the graph-encoded prior expectation is chosen as the actual network structure;

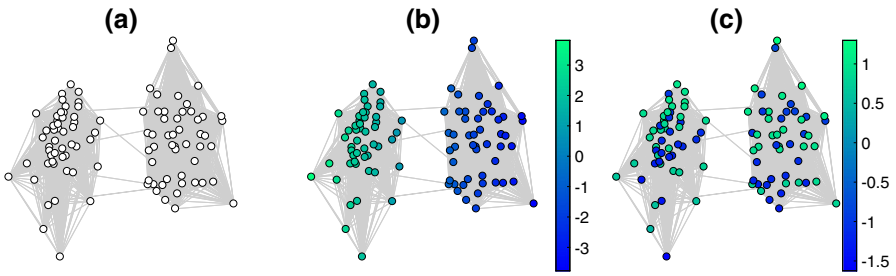

Fig. 2 Communities data (Sect. 3.3), **a** the actual network, **b** nodes colored according to their projected values using the first PCA component **c** similar to (**b**), but for the first SICA component (our method). The x-axis corresponds to the first feature in the data, while the position of points on the y-axis is picked at random. The PCA projection picks up the variance across the clusters, while the SICA projection highlights the variance within the clusters

Table 1 Communities data (Sect. 3.3), weights of first component for PCA and SICA

	Feature 1	Feature 2	···
PCA 1st component	− 0.998	0.015	···
SICA 1st component	0.186	0.957	···

i.e., the resulting prior graph consists of the two cliques and a few edges in-between, see Fig. 2a.

We compare the primary projections given by PCA and SICA, see Fig. 2b, c. For both the PCA and SICA projections, we colored the data points according to their projected values, i.e., $\hat{\mathbf{X}}\mathbf{w}$, where $\mathbf{w}$ correspond to the first component of PCA/SICA. In Fig. 2b, we see that the PCA projection gives one cluster a higher score (green vertices) than the other (blue vertices). Clearly, PCA picks up the structure of the two communities defined by the first attribute. In contrast, SICA assigns both high and low scores within each cluster (Fig. 2c). That is, it highlights variance *within* the clusters. This is to be expected, because the community structure is very similar to the graph structure, with which we assume the user knows already.

Table 1 lists the weight vectors of the projections. As expected, PCA gives a large weight to the first feature, which has higher variance. However, SICA's first component is dominated by the second feature. Hence, by incorporating the community structure as prior expectation, SICA finds an alternative structure corresponding to the second feature.

3.4 Discussion: potential and limitations of SICA

Potential of SICA The three instantiations of SICA discussed in this section are illustrative of SICA's potential to take into account prior beliefs of the data analyst, and to find projections that are interesting with respect to it. The three steps that need to be followed to instantiate SICA are always the same: (1) Express the prior belief in the form of constraints on the expected value of certain specified statistics—i.e. in form of Eq. (1)—and solve the MaxEnt problem (9) to obtain the background distribution.

(2) Compute the marginal density function of the background distribution for the data projection onto a projection matrix $\mathbf{W}$. And (3), come up a good optimization strategy. *In principle*, any data analyst able to express their prior beliefs in the required form can thus benefit from this approach.

Limitations of SICA Yet, each of these steps also implies some important limitations of SICA that should be the subjects of further work. The result of the first step will always be an exponential family distribution, and hence have an analytical form. However, expressing prior belief types as required will often be beyond the capabilities of a data analyst. Also the second step may require considerable mathematical expertise. Indeed, it may not be possible to express the marginal distribution in an analytical form such that it may need to be approximated. And even when it can be expressed analytically, deriving it mathematically may be non-trivial. Finally, thanks to the orthonormality assumption of the projection matrix, general purpose (Stiefel) manifold optimization solvers are in principle applicable, but doing this does not provide any optimality guarantees.

SICA in practice For these reasons, SICA as a framework is not directly suitable for use by practitioners. Instead, it can be used by researchers to develop specific instantiations of sufficiently broad applicability, which can then be made available to practitioners. Probably the most powerful example of this is the third instantiation (Sect. 3.3). Indeed, it is a very generic prior belief type for which an efficient algorithm exists, and which is relatively easy to be used.

4 Experiments

In this section, we present several case studies which demonstrate how SICA may help users to explore various types of real world data. For every case, we specify some background knowledge a user might have, and encode that knowledge using previously defined expressions. The encoded beliefs are then provided to SICA in the form of the background distribution. Third, we analyze the projections computed by SICA and evaluate whether they are indeed interesting with respect to the assumed user's prior. Finally, we summarize the runtime of all experiments presented in this section.

Note that the purpose of our experiments is not to investigate superiority of SICA over existing methods for dimensionality reduction. Instead, we aim to investigate whether and to which extent SICA's results usefully depend on the various prior beliefs, in highlighting information that is complementary to them. Where the answer to this question is positive, SICA is the method of choice—of course, assuming the prior beliefs are well-specified.

4.1 t-PCA on real-world data

Setup We evaluate the use of SICA with a spread prior (t-PCA) on two datasets. The Shuttle[6] data describes radiator positions (seven position classes: Rad Flow, Fpv

[6] https://archive.ics.uci.edu/ml/datasets/Statlog+(Shuttle), retrieved November 18, 2016.

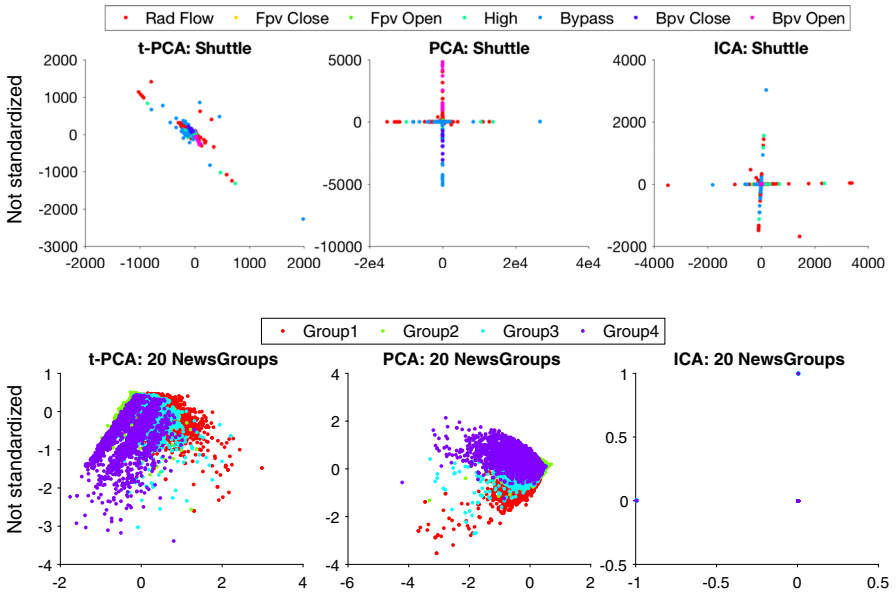

Fig. 3 Real world data case study for t-PCA (Sect. 4.1). The top 2 projections found by t-PCA (left), PCA (middle), and FastICA (right). Top row: Shuttle; bottom row: 20 NewsGroups. For the Shuttle dataset, the PCA and FastICA projections show highest variances as well as the most independent dimensions. SICA projection exhibits other, smaller-scale variation. For the 20 NewsGroup dataset, SICA's result is qualitatively similar to PCA's result but with slightly lower variance. The FastICA's result is qualitatively different

Close, Fpv Open, High, Bypass, Bpv Close) in a NASA space shuttle and consists of 58000 data points and 9 integer attributes, i.e., $\hat{\mathbf{X}} \in \mathbb{R}^{58000\times 9}$. The 20 Newsgroups[7] data describes four newsgroups (four classes) and has 16242 points and 100 integer attributes, i.e., $\hat{\mathbf{X}} \in \mathbb{R}^{16242\times 100}$. Both datasets are centered such that each attribute has zero mean.

Both of the datasets contain complex structures. Particularly, the shuttle dataset contains highly imbalanced cluster structure: one of the classes forms 80% of the population. For both datasets, we assume the user has a prior belief only about the order of magnitude of the data, i.e., the user would not be surprised by the presence of outliers. This can be encoded using the spread prior with a small ρ, e.g., $\rho = 10^{-5} \cdot (\frac{1}{|\mathbf{X}|}\sum_{i=1}^{|\mathbf{X}|} ||x_i||_2)^{\frac{1}{2}}$.

Results We compared the results of SICA, PCA, and FastICA[8] (Hyvärinen 1999). FastICA is a popular PP method that implements ICA. We used FastICA with default parameters. The classes for each dataset are plotted in different colors.

Figure 3 shows the results of SICA with this prior belief model, for PCA, and for FastICA. For the Shuttle dataset, PCA and FastICA give visually similar results: the

[7] http://cs.nyu.edu/~roweis/data.html, retrieved November 18, 2016.

[8] In the experiment we used the FastICA package for MATLAB. The package can be downloaded from https://research.ics.aalto.fi/ica/fastica/

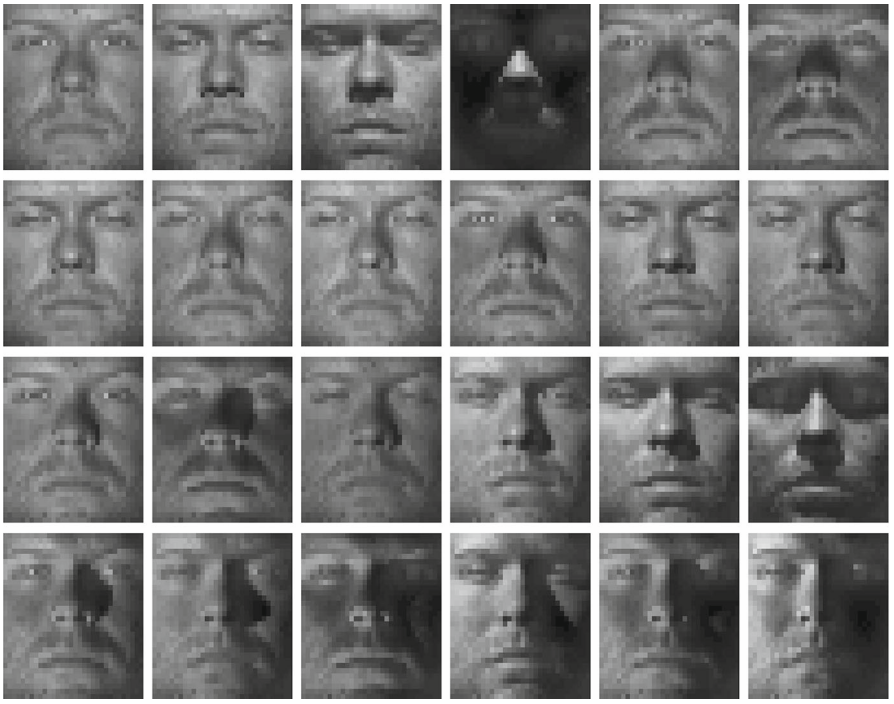

Fig. 4 Faces dataset (Sect. 4.2), subject one, first 24 lighting conditions. The data set contains 31 human subjects where each of them has face image taken under 64 lighting conditions. Each face image contains 32×32 pixels

highest-variance as well as the most independent dimensions appear to be affected by relatively few data points with large projection values along them. Especially for PCA, the resulting scatter plot has axes with very large scales. Hence the data points that correspond to small scale structure are more likely to be plotted on top of each other, making them harder to discern. SICA, in accounting for order of magnitude variations in the norms of data points, is less biased towards these distant data points. As a result, it prefers lower-variance projections which exhibit other, smaller-scale variation, which therefore provide information that complements the user's expectations.

For the 20 Newsgroup dataset, SICA's result is qualitatively similar to PCA's result, although the variance of the SICA projection is slightly lower, arguably in favor of making the more fine-grained variation in the data more apparent. FastICA's result, however, is qualitatively different. It puts all weight on a single binary attribute, such that its top components project all data points onto just three points.

4.2 Images and lighting, with a graph prior

Setup We now apply SICA to explore image data. The Extended Yale Face Database B[9] contains frontal images of 38 human subjects under 64 illumination conditions,

[9] This data is available as a preprocessed Matlab file at http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html. The original dataset is described in Georghiades et al. (2001), Lee et al. (2005).

for example, see Fig. 4. We ignored the images of seven subjects whose illumination conditions are not fully specified. The input dataset then contains 1684 data points, each of which is described by 1024 real valued features , i.e., $\hat{\mathbf{X}} \in \mathbb{R}^{1684\times 1024}$. The data is then centered to have a zero mean. The task of decomposing images in order to account for a number of pre-specified factors has been addressed in the past (e.g., using a N-mode SVD; Vasilescu and Terzopoulos 2002). Here we want to explore how SICA weight vectors change according to the prior belief of a specific user.

Let us assume that the user already knows there are lighting conditions and is not interested in them. We can encode such knowledge by declaring that images (data points) with the same lighting condition are similar to each other. This can be expressed in a point-wise similarity prior. We construct a graph where each image is a node and two nodes are connected by an edge if the corresponding images have the same lighting conditions. The resulting prior graph consists of 64 cliques, one for each lighting condition.

Results We compare the weight vectors of the subjectively interesting components (SICs) given by SICA and top principal components (PCs) given by PCA, namely the Eigenfaces from the two methods. We expect PCA to find a mixture of illumination and facial features, while SICA should find mainly facial structure. Note that illumination conditions vary similarly across the human subjects, while facial structures are *subject specific*. The principal Eigenfaces from PCA and SICA are presented in Fig. 5. We observe that the Eigenfaces given by PCA are influenced substantially by the variation in lighting conditions. These conditions vary from back-to-front, right-to-left, top-to-down, down-to-top and left-top-to-right-bottom. Because the images of each subject contain every lighting condition, it appears indeed more difficult to separate the subjects based only on the top PCA components. On the other hand, the Eigenfaces from SICA highlight local facial structures, like the eye area (first, third and fifth faces), and the mouth and nose (first, third and fifth faces). Note though that the first and second SICA Eigenfaces also still pick up some lighting variation, which is confirmed by the similarity between the top two SICA and PCA components (left upper corner of Fig. 6). The absolute value of the inner product between the first SICA and PCA components is 0.91 and the value of the second components is 0.93. Note also that the similarities of most other SICA and PCA components are considerably smaller, confirming that SICA components are indeed truly different from the PCA components.

If SICA succeeds in providing insights that contrasts with the prior beliefs about the lighting conditions, the projection of an image onto the top SICs can be expected to separate the subjects better than the projection onto an equal number of top PCs. To verify this, we computed the 10-fold cross-validation loss (with respect to the subjects as labels) of a k-Nearest Neighbors (k-NN) classifier on the projected features with respect to the top PCs and SICs. A projection that separates the subjects well will have low classification loss. We applied k-NN on the SICA/PCA projections with number of components ranging from 1 to 50. Since our goal here is to evaluate whether top SICs are more likely to correspond to facial structure than top PCs, rather than achieve best classification accuracy, we fix $k = 3$. Figure 7a shows that indeed top SICs (orange line) give a better separation than top PCs (purple line). In addition, we performed the same experiment using an SVM (rather than 3-NN) with 10-fold cross validation

Fig. 5 Faces data case study (Sect. 4.2), top five Eigenfaces for PCA (top) and SICA (bottom). The Eigenfaces from PCA are influenced substantially by the variation in lighting conditions, while the Eigenfaces from SICA mainly highlight local facial structures

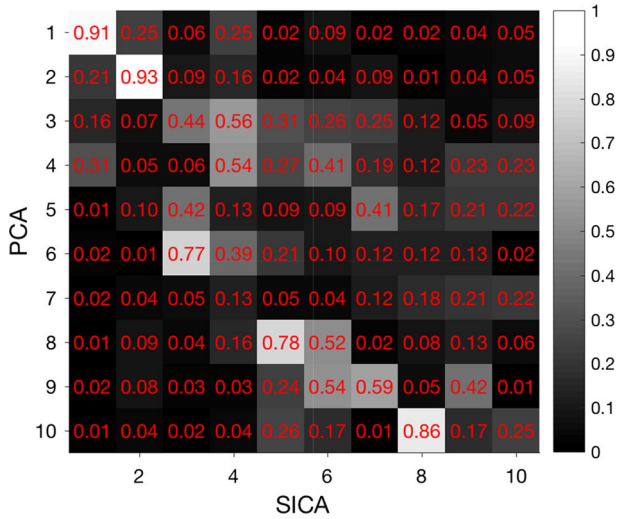

Fig. 6 Face data case study (Sect. 4.2), Similarity (absolute value of inner product) between PCA and SICA top 10 components. The similarity between the top two SICA and PCA components confirms that SICA top two components still pick up some lighting variation. The less significant similarity between the other SICA and PCA components indicates SICA components are indeed truly different from the PCA components

on the projected features to perform classification. We measured the average losses over 10 folds while varying the number of projected features form 1 to 50. The result (Fig. 8a) shows SICA is more accurate than PCA when the number of features is small. PCA then catches up when the number of the dimensions increases.

Conversely, as SICA with the stated prior beliefs should result in a projection that highlights information *complementary* to lighting conditions, one can expect that the

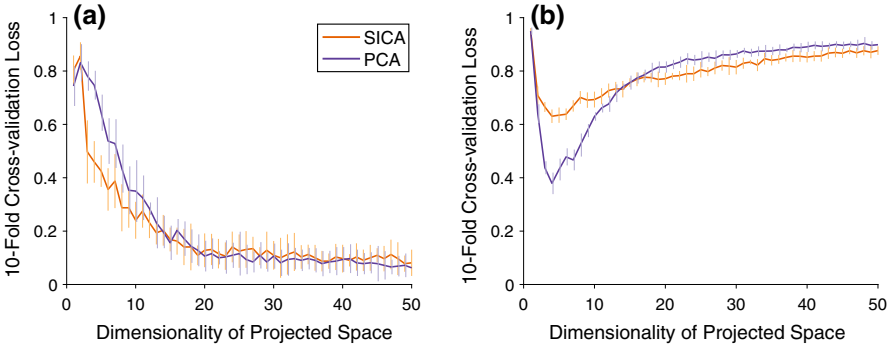

Fig. 7 Faces data case study (Sect. 4.2), **a** average 10-fold cross-validation loss (error bars gives one standard deviation, the smaller loss the better) for 3-NN subject classification on the projected data. Top SICs gives better separation of subjects than top PCs. **b** Average 10-fold cross-validation loss (error bars gives one standard deviation, the smaller loss the better) for 3-NN lighting condition classification. Top PCs gives better separation of lighting conditions than SICs

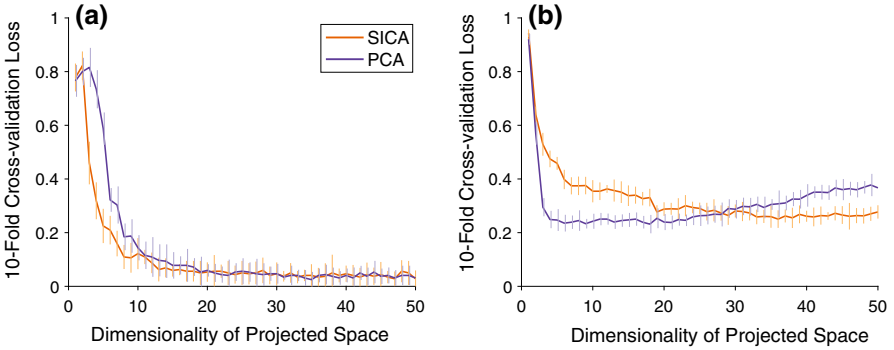

Fig. 8 Faces data case study (Sect. 4.2), **a** average 10-fold cross-validation loss (error bars gives one standard deviation, the smaller loss the better) for SVM subject classification on the projected data. Top SICs gives better separation of subjects than top PCs. **b** Average 10-fold cross-validation loss (error bars gives one standard deviation, the smaller loss the better) for SVM lighting condition classification. Top PCs gives better separation of lighting conditions than SICs

top SICs perform worse in separating the different lighting conditions than the top PCs. To evaluate this, instead of classifying subjects, we used k-NN to classify different illumination conditions, using the same PCs and SICs as before. That is, where we told SICA explicitly we were not interested in light variation. Figure 7b shows that PCA indeed gives better 3-NN classification accuracy than SICA. The result (Fig. 8b) obtained by SVM confirms this with another classifier.

4.3 Spatial socio-economy, with a graph prior

Now we use SICA to analyze a socio-economic dataset. The German socio-economic data (Boley et al. 2013) was compiled from the database of the German Federal Statistical Office. The dataset consists of socio-economic records of 412 administrative

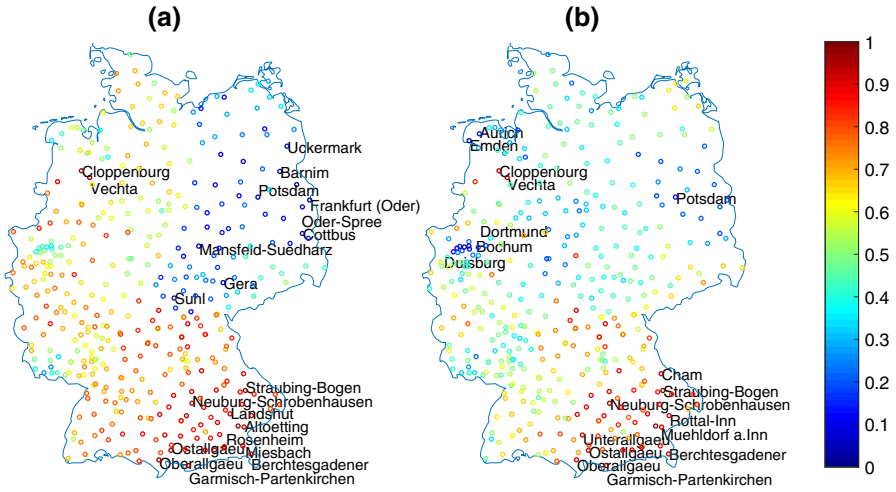

Fig. 9 German socio-economics data vote attributes (Sect. 4.3.1). **a** The geographic scatter plot of districts with each district colored according to its projective value onto top PC. The top 10 districts with most positive and most negative projective values are labeled. The top PC assigns low scores to the districts in East Germany, while it gives rest districts relatively high scores. **b** The same geographic scatter plot for the top SIC. Although SICA still shows considerable global variation (in this case between the north and the south), it also highlights the variations that are more local

districts in Germany. The data features used in this case study fall into two groups: election vote counts and age demographics. We additionally coded for each district the geographic coordinates of the district center and which districts share a border with each other.

4.3.1 Vote attribute group

Setup Let us assume a user is interested in exploring the voting behavior of different districts in Germany. The (real-valued) data attributes about the 2009 German elections cover the percentage of votes on the five largest political parties[10]: CDU/CSU, SPD, FDP, GREEN, and LEFT. Thus, we have a dataset $\hat{\mathbf{X}} \in \mathbb{R}^{412\times 5}$. We centered the data attribute-wise by subtracting the mean from each data point.

Let us assume also that the user already knows the East–West divide has still a large influence. Hence, she may believe the voting behavior of the districts in the east are similar to each other, and the same goes for the west. This prior belief can also be encoded as point-wise similarities. By treating each district as a graph node, we can translate our knowledge into prior expectations, by connecting similar districts with edges. This results in a graph with two cliques: one clique consists of all districts in East Germany, the other clique contains the rest.

Results The projection onto the first PC (Fig. 9a) shows smooth variation across the map. Districts in western Germany and Bavaria (south) receive high scores (red circles)

[10] https://en.wikipedia.org/wiki/List_of_political_parties_in_Germany, retrieved November 18, 2016.

Table 2 German socio-economics data vote attributes (Sect. 4.3.1), weights given by top PCA and SICA component

	CDU/CSU	SPD	FDP	GREEN	Left
PCA 1st	0.53	−0.13	0.22	0.13	−0.80
SICA 1st	0.72	−0.65	0.10	−0.09	−0.19

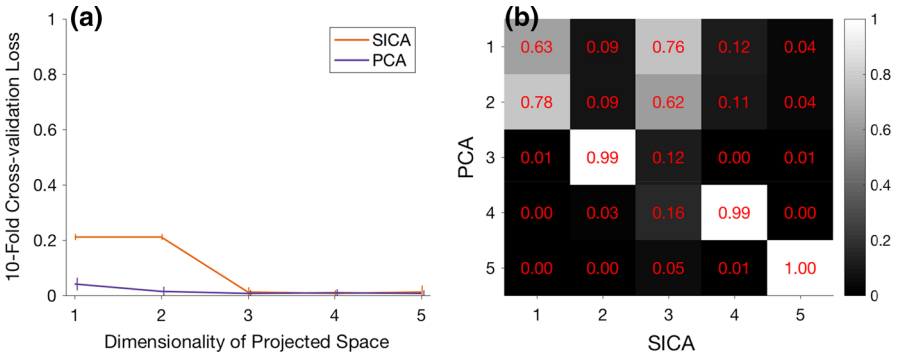

Fig. 10 German socio-economics data vote attributes (Sect. 4.3.1). **a** Average 10-fold cross-validation loss (error bars gives one standard deviation) for eastern and non-eastern districts classification on the projected data. The top two PCs lead to a smaller loss than the top two SICs. **b** Similarity (absolute value of inner product) between PCA and SICA components. The first and second components of the two methods are different. The third SIC is similar to the first and second PCs

and districts in East Germany (Brandenburg and Thuringa) have low scores (dark blue circles). Table 2 additionally shows the weight vectors of the top PC and SIC. The PC is dominated by the difference between CDU/CSU and Left. This is expected, because this indeed is the primary division in the elections; East Germany votes more Left, while in Bavaria, CSU is very popular.

However, SICA highlights a different pattern; the competition between CDU/CSU and SPD is more local. Although there is still considerable global variation (in this case between the south and the north), we also observe that the Ruhr area (Dortmund and around) is similar to East Germany in that the social-democrats are preferred over the Christian parties. Arguably, the districts where this happens are those with a large fraction of working class, like the Ruhr area. Perhaps they vote more on parties that put more emphasis on interests of the less-wealthy part of the population.

To investigate this in a more quantitative manner, we applied an SVM to classify the eastern versus non-eastern districts using the vote data projected onto the top SICA and PCA components. We measured the 10-fold cross-validation losses for the projected data's dimensionality ranging from 1 to 5. Figure 10a shows that the first two PCA components lead to a smaller loss than SICA. This indicates that the two top PCs indeed reflect more to the eastern and non-eastern division. The similarity matrix (Fig. 10b) of the PCs and SICs also shows that the first and second components of the two methods are different. Notice that the third SIC (third column) is similar to the first and second PCs. This explains why when the dimensionality of projected space increased to three, the classification loss of SICA drops to the same as PCA.

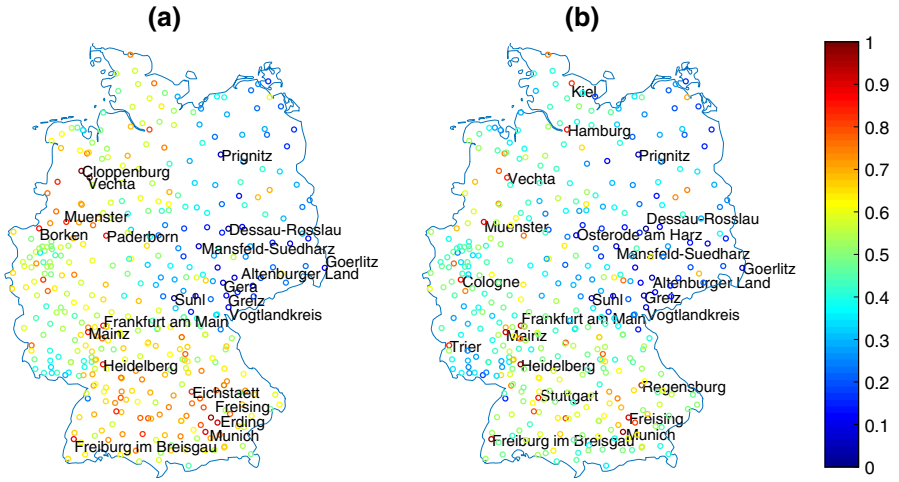

Fig. 11 German socio-economics data demographic attributes (Sect. 4.3.2). **a** The geographic scatter plot of districts with each district colored according to its projective value onto first PC. The top 10 districts with most positive and most negative projective values are labeled. The PC again highlights the difference between East and West Germany. **b** The same geographic scatter plot against first SICA component. The top SIC assigns large negative scores to East Germany, while it also highlights the large cities

4.3.2 Demographic attribute group

Setup Next, we assume that the user is interested in exploring the age demographics of different districts. The demographic attribute group describes the age distribution of the population (in fractions) for every district, over five categories: *Elderly* (age > 64), *Old* (between 45 and 64), *Middle Aged* (between 25 and 44), *Young* (between 18 and 24), and *Children* (age < 18), represented by a positive real-valued vector of length 5. Thus, we have a data set $\hat{\mathbf{X}} \in \mathbb{R}^{412\times 5}$. We then centered the data attribute-wise.

We assume again the user understands the influence of the historical east–west divide. We are interested in finding patterns orthogonal to that division. The population density is lower in East Germany than the rest of country. According to Wikipedia[11]: "About 1.7 million people have left the new federal states since the fall of the Berlin Wall, or 12% of the population. A disproportionately high number of them were women under 35". Also the Berlin-Institute for Population and Development[12] reports: "the birth rate in East Germany dropped down to 0.77 after unification, and raised to 1.30 nowadays compare to 1.37 in the West". Given this (in Germany common sense) knowledge, SICA should be able to offer new insights. Hence, we assume again that the demographics of the districts in East Germany are similar, and the remaining districts are also similar. Formalizing such belief as point wise similarities results in a

[11] https://en.wikipedia.org/wiki/New_states_of_Germany#Demographic_development, retrieved November 18, 2016.

[12] http://www.berlin-institut.org/fileadmin/user_upload/Studien/Kurzfassung_demografische_lage_englisch.pdf

Table 3 German socio-economics data age demographics (Sect. 4.3.1), weights given by first PCA and SICA component

	Elderly	Old	Mid-age	Young	Child
PCA	−0.61	−0.42	0.43	0.09	0.51
SICA	−0.62	−0.32	0.69	0.19	0.06

graph with two cliques: one consists of all districts in East Germany, the other contains the rest.

Results Projection on the top PC (Fig. 11a) confirms the user's prior expectations. There is a substantial difference between East and West Germany. In the visualization, high projection values (red color) appear mostly in East Germany, while low values (blue color) appear mostly in the rest of Germany. If we look at the weights of the top PC (Table 3), we find that the projection is based on large negative weights to people above 44 (Old and Elder), and large positive weights to the younger population (age < 45). This confirms that indeed the demographic status of East Germany deviates.

SICA results in a different projection (Fig. 11b), even though the difference is more subtle than in the analysis of the voting behavior. Although SICA also assigns large negative scores to East Germany, presumably because there are relatively many elderly there, SICA also highlights the large cities, e.g., Munich, Cologne, Frankfurt, Hamburg, Kiel, Trier. In addition to showing a smooth geographic East–West trend, SICA also seems to highlight districts whose demographic status deviates from its surrounding districts. Indeed, from the weight vector (Table 3) we see that these districts are found by considering the number of middle aged people against the number of elderly. We know that many middle-aged (24–44) working people live in large cities, and, according to the report from Berlin-Institute for Population and Development, "large cities generally have fewer children, since they offer families too little room for development". Indeed, we find that families live in the neighboring districts, highlighting a perhaps less-expected local contrast.

Also, to further investigate this more quantitatively, we applied an SVM to classify the eastern versus non-eastern districts using the projected demographic attributes. Figure 12a shows that the top two PCA components result in a slightly smaller loss than SICA. This indicates that the top PCs and SICs both correspond to the eastern and non-eastern division. The similarity matrix (Fig. 12b) of PCA and SICA components also shows the first and second components of the two methods are very similar. However, according to the visualization, the best (positively) scored districts by SICA (Fig. 11b) highlight large cities more than the PCA result (Fig. 11a). Also the highlighted cities stand out more from their surrounding area.

4.4 Runtime

Table 4 summarizes the runtime of PCA and SICA in all experiments presented in this paper. In all these cases, SICA takes more time to compute the projections. For the first three columns (t-PCA cases), we used the solver offered by Manopt to perform gradient descent over the Stiefel manifold. We tried ten random starts in all three cases and picked the projection that gives the best objective. The ten random starts already give stable local optima in all three cases. Note that t-PCA scales gracefully when the

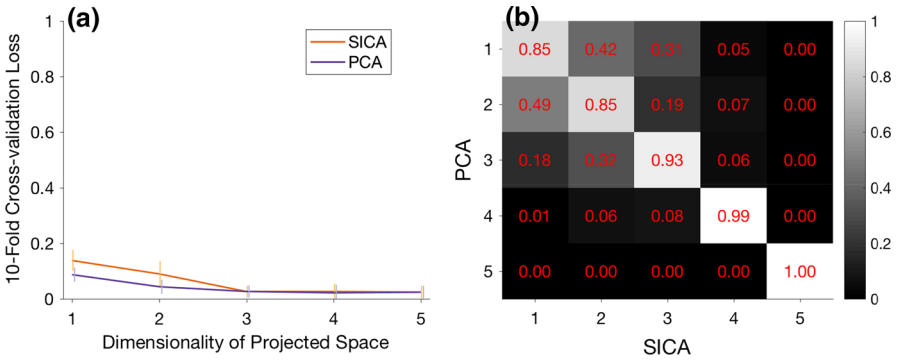

Fig. 12 German socio-economics data age demographics (Sect. 4.3.2). **a** Average 10-fold cross-validation loss (error bars gives one standard deviation) for eastern and non-eastern districts classification on the projected data. The top two PCA components result in a slightly smaller loss than SICA. **b** Similarity (absolute value of inner product) between PCA and SICA components. The first and second components of the two methods are very similar

Table 4 Runtime (in seconds) of SICA and PCA for all experiments (Sect. 4.4)

	Synthetic outlier	Shuttle	20NewsGroup	Synthetic community	Socio-eco. (age)	Socio-eco. (vote)	Face image
SICA	0.12	1.75	8.07	0.03	0.06	0.04	2.26
PCA	< 0.01	0.08	0.25	0.01	< 0.01	< 0.01	0.56

Each measurement is averaged over ten runs. We used a machine with Intel Quad Core 2.7 GHz CPU and 16 GB 1600 MHz DDR3 RAM

data size increases from Synthetic dataset (1010×2) to Shuttle (58000×9) and then 20NewsGroup (16242×100).

The other four experiments are about SICA with graph prior. Again, SICA scales well from the Synthetic data (100×10) to the socio-economical dataset (both 412×5) and then the Face image dataset (1684×1024). However, although both SICA and PCA are based on eigenvalue decomposition, SICA spend more time than PCA. One reason is that in order to construct the Laplacian matrix in (40) SICA needs to loop through the data as well as find the best multipliers. Note that the current experiments are based on a quick implementation—a more careful implementation may improve the run time of SICA.

5 Related work

SICA is linear, unsupervised, and subjective. Dimensionality reduction (DR) methods, as indicated by the name, aim to find lower dimensional representation of high dimensional data. Here "dimension" refers to the number of features that are used to describe the data. Finding a lower dimensional representation further boils down to either select a subset of the original features or transform the feature space to another (low-dimensional) space. Here we mainly discuss the line of work for feature transformation (extraction), since they are more closely related to our work.

Supervised vs. unsupervised DR methods are often designed with a certain goal: to have lower dimensional representations with some specific properties. For example Principal Component Analysis (PCA) (Peason 1901; Jolliffe 2002) is often used for computing a presentation of dataset where the data variance is preserved, whereas Canonical Component Analysis (CCA) (Hotelling 1936) aims to find pairs of directions in two feature spaces where the corresponding two datasets are highly correlated. While PCA and CCA achieve their goals in an unsupervised manner, Linear Discriminant Analysis (LDA) (Fisher 1936), on the other hand, extracts discriminative features according to the given class labels with a supervised flavor. The new features provided by DR methods can not only be used for later classification or prediction, but also to explore the structures in the data, e.g., Self Organizing Map (SOM) (Kohonen 1998) for exploratory data analysis. In order to meet different analysis goals under a unified framework, Projection Pursuit (PP) (Friedman and Tukey 1974) was proposed to locate different projections according to some predefined "interestingness index". Different from the previous works, we seek for data projections that are interesting particularly to the user. Therefore, SICA aims to propose a generic interestingness measure that does not explicitly depend on the context of the data or on the specific analytic tasks.

Linear vs. non-linear Orthogonally, when approaching these goals, DR methods further assume the relationship between the original data and its lower dimensional representation to be either linear or non-linear. The aforementioned methods (PCA, CCA and LDA) compute new data representations via linear transformation. Additionally, classical Multidimensional Scaling (Kruskal and Wish 1978) also finds a linear transformation that preservers the distances between the data points. We refer the reader to the survey by Cunningham and Ghahramani (2015) and the references therein for a comprehensive review of linear DR techniques. However, in reality, high dimensional data often obeys certain constraints; data then lies on a low-dimensional (non-linear) manifold embedded in the original feature space. Non-linear dimensionality reduction methods like SOM approximate such a manifold by a set of linked nodes. Building upon Multidimensional Scaling, ISOMAP (Tenenbaum et al. 2000) seeks to preserve the intrinsic geometry of the data by first encoding neighborhood relations as a weighted graph. This inspired later spectral methods (Von Luxburg 2007; Ng et al. 2002) as well as different manifold learning approaches (Belkin and Niyogi 2003; He and Niyogi 2004; Weinberger et al. 2006) that try to solve a eigenproblem in order to discover the intrinsic manifold structure of the data, using an eigendecomposition to preserve the local properties of the data. Note that with a graph prior, SICA computes linear projections in a spectral-method-like manner (Sect. 3.3). However, the previously mentioned non-linear DR methods are interested in the eigenvectors corresponding to the smallest k eigenvalues of the Laplacian, as they provide insights into the local structure of the underlying graph, while SICA identifies mappings that *target* non-smoothness with respect to the user's beliefs about the data, while maximizing the variance of the data in the resulting subspace. Interestingly, the resulting optimization problem is not simply the opposite of existing approaches.

Objective vs. subjective The aforementioned methods are mainly "objective" in the sense there that user is not explicitly considered. A notable exception is the work on User Intent Modeling for Information Discovery (Ruotsalo et al. 2015), where indeed

an explicit relevance model is built to help a user find information relevant to her search. Their tool also computes a 2D embedding of the search results, accounting for their user and session specific relevance. However, they do not introduce a new theoretically well-motived method to find a low-dimensional subspace that accounts for background knowledge or intent. That is also not the focus of their work, which is rather the identification of relevant results. Some other techniques have been proposed in exploratory data analysis that take into account the user knowledge to determine interesting projections. For instance, Brown et al. (2012) suggests an interactive process in which the user provides feedback by moving incorrectly-positioned data points to locations that reflect their understanding. In a similar manner, Paurat and Gärtner (2013) make use semi-supervised least squares projections but allowing the user to select and rearrange some of the embedded data points. In the work by Iwata et al. (2013), the authors use active learning to select candidate data points for the user to relocate so that they can achieve their desired visualization. All of these methods, guided by the user, interactively present different aspects of the data. Finally the work by Weinberger and Saul (2009) require the practitioner to provide auxiliary information, e.g. a similarity graph, that identify target neighbours for each data point, that is then used to constraint their optimization problem. This prior knowledge is the structure that one wants to preserve, as opposed to SICA. To our best knowledge, SICA is the first subjective DR method which finds lower-dimensional data representations that are as interesting as possible for a particular user. Hence, SICA adds another layer to the family of dimensionality reduction methods.

6 Conclusion

In exploratory data analysis, structures in the data often have different value for different tasks and data analysts. To address this, the Projection Pursuit literature has introduced numerous *projection indices* that quantify the interestingness of a projection in various ways. However, it still seems to be conceptually challenging to define a generic quality metric for the tasks of exploratory data analysis. As an attempt in this direction, we present SICA, a new linear dimensionality reduction approach that explicitly embraces the subjective nature of interestingness. In this paper, we show how the modeling of a user's belief state can be used to drive a subjective interestingness measure for DR. Such interestingness measure is then used to search for subjectively interesting projections of data. Results from several case study show that it can be meaningful to account for available prior knowledge about the data.

Avenues for further work include incorporating multiple prior expectations simultaneously (e.g., define multiple (disjoint) groups of similar nodes using graph prior), to enable more flexible iterative analysis. This involves solving a MaxEnt optimization problem subject to multiple constraints. We also plan to study how to improve the interpretability of the projections, e.g., finding projections with sparse weight vectors. In terms of visualization, an interesting future direction is to investigate how the SICA result will be affected by removing the assumption of the resolution being the same through all dimensions. Although that is already possible, one question is how a user could conveniently input these expectations into the system. Another open question is

to what extend SICA can be applied to non-linear dimensionality reduction. Finally, alternative types of prior expectations are also worth examining.

Acknowledgements We thank the anonymous reviewers for their constructive and insightful comments. We are grateful to Petteri Kaski for discussions about the complexity of t-PCA.

Appendix A Probability approximation based on distribution (17)

We want to show that given marginal density function (17) the probability $\Pr(\mathbf{XW} \in [\hat{\mathbf{\Pi}}_{\mathbf{W}} - \Delta\mathbf{1}, \hat{\mathbf{\Pi}}_{\mathbf{W}} + \Delta\mathbf{1}])$ can be approximated well by using the form $p_{\mathbf{\Pi}_{\mathbf{W}}}(\mathbf{XW}) \cdot 2\Delta$ for sufficiently small Δ. As random variable $\mathbf{XW}$ in distribution (17) consists of elements that are all independent to each other, it is sufficient to show the approximation quality for one dimensional normal distribution $\mathbf{x} \sim \mathcal{N}(0, \sigma^2)$:

Proposition 1 *For one dimensional normal random variable* $\mathbf{x} \sim \mathcal{N}(0, \sigma^2)$*, the approximation of probability* $Pr(\mathbf{x} \in [\mathbf{x} - \Delta, \mathbf{x} + \Delta])$ *by* $p_{\mathbf{x}}(\mathbf{x}) \cdot 2\Delta$ *has a bounded log approximation ratio:*

$$\left| \log \left(\frac{Pr(\mathbf{x} \in [\mathbf{x} - \Delta, \mathbf{x} + \Delta])}{p_{\mathbf{x}}(\mathbf{x}) \cdot 2\Delta} \right) \right| \leq \begin{cases} \frac{\Delta(2|\mathbf{x}| + \Delta)}{2\sigma^2} & : |\mathbf{x}| \geq \Delta, \\ \frac{3\Delta^2}{2\sigma^2} & : |\mathbf{x}| \leq \Delta. \end{cases}$$

Thus, for given σ and $\mathbf{x}$, if Δ is sufficiently small and $\mathbf{x}\Delta$ tends to 0, the upper bound of the log approximation ratio tends to zero. Namely, the approximation is tight.

Proof Let us first consider the case where $\mathbf{x} - \Delta > 0$. Because of the symmetry of the normal distribution, the result also applies for the case where $\mathbf{x} + \Delta < 0$.We have:

- Estimation of the probability: $2\Delta \cdot 1/\sqrt{2\pi\sigma^2} e^{-\mathbf{x}^2/(2\sigma^2)}$.
- Upper bound on the probability: $2\Delta \cdot 1/\sqrt{2\pi\sigma^2} e^{-(\mathbf{x}-\Delta)^2/(2\sigma^2)}$.
- Lower bound on the probability: $2\Delta \cdot 1/\sqrt{2\pi\sigma^2} e^{-(\mathbf{x}+\Delta)^2/(2\sigma^2)}$.

Then the log approximation ratio can be computed as:

1. for upper bound we have

$$\left| \log \left(\frac{2\Delta \cdot 1/\sqrt{2\pi\sigma^2} e^{-(\mathbf{x}-\Delta)^2/(2\sigma^2)}}{2\Delta \cdot 1/\sqrt{2\pi\sigma^2} e^{-\mathbf{x}^2/(2\sigma^2)}} \right) \right| = \left| \log \left(\frac{e^{-(\mathbf{x}-\Delta)^2/(2\sigma^2)}}{e^{-\mathbf{x}^2/(2\sigma^2)}} \right) \right| = \frac{\Delta(2\mathbf{x} - \Delta)}{2\sigma^2}$$

2. and for lower bound we have

$$\left| \log \left(\frac{2\Delta \cdot 1/\sqrt{2\pi\sigma^2} e^{-(\mathbf{x}+\Delta)^2/(2\sigma^2)}}{2\Delta \cdot 1/\sqrt{2\pi\sigma^2} e^{-\mathbf{x}^2/(2\sigma^2)}} \right) \right| = \left| \log \left(\frac{e^{-(\mathbf{x}+\Delta)^2/(2\sigma^2)}}{e^{-\mathbf{x}^2/(2\sigma^2)}} \right) \right|$$

$$= \frac{\Delta(2\mathbf{x} + \Delta)}{2\sigma^2}$$

Since $\mathbf{x}, \Delta > 0$, the absolute log approximation ratio of lower bound is always smaller than the ratio achieved by the upper bound, we have for $\mathbf{x} - \Delta > 0$:

$$\left| \log \left(\frac{p_{\mathbf{x}} \left(\mathbf{x} \in (\mathbf{x} - \Delta, \mathbf{x} + \Delta) \right)}{p_{\mathbf{x}}(\mathbf{x}) \cdot 2\Delta} \right) \right| \leq \frac{\Delta(2\mathbf{x} + \Delta)}{2\sigma^2} \tag{41}$$

Given σ and $\mathbf{x}$, if Δ is sufficiently small such that $\mathbf{x}\Delta$ close to 0, then the approximation at $\mathbf{x}$ ($|\mathbf{x}| \geq \Delta$) is tight.

Remark 6 In general, for $|\mathbf{x}| \geq \Delta$, the right hand side in inequality (41) can be replaced by $\frac{\Delta(2|\mathbf{x}|+\Delta)}{2\sigma^2}$

Let us now consider the case where $-\Delta \leq \mathbf{x} \leq \Delta$. Without losing generality, we assume $p(\mathbf{x} - \Delta) > p(\mathbf{x} + \Delta)$. This leads to:

- Estimation of the probability: $2\Delta \cdot 1/\sqrt{2\pi\sigma^2} e^{-\mathbf{x}^2/(2\sigma^2)}$.
- Upper bound on the probability: $2\Delta \cdot 1/\sqrt{2\pi\sigma^2}$.
- Lower bound on the probability: $2\Delta \cdot 1/\sqrt{2\pi\sigma^2} e^{-(\mathbf{x}+\Delta)^2/(2\sigma^2)}$.

Then the log approximation ratio can be computed as:

1. for upper bound we have

$$\left| \log \left(\frac{2\Delta \cdot 1/\sqrt{2\pi\sigma^2}}{2\Delta \cdot 1/\sqrt{2\pi\sigma^2} e^{-\mathbf{x}^2/(2\sigma^2)}} \right) \right| = \left| \log \left(\frac{1}{e^{-\mathbf{x}^2/(2\sigma^2)}} \right) \right| \tag{42}$$

$$= \frac{\mathbf{x}^2}{2\sigma^2} \tag{43}$$

$$\leq \frac{\Delta^2}{2\sigma^2}, \tag{44}$$

2. and for lower bound we have

$$\left| \log \left(\frac{2\Delta \cdot 1/\sqrt{2\pi\sigma^2} e^{-(\mathbf{x}+\Delta)^2/(2\sigma^2)}}{2\Delta \cdot 1/\sqrt{2\pi\sigma^2} e^{-\mathbf{x}^2/(2\sigma^2)}} \right) \right| = \left| \log \left(\frac{e^{-(\mathbf{x}+\Delta)^2/(2\sigma^2)}}{e^{-\mathbf{x}^2/(2\sigma^2)}} \right) \right| \tag{45}$$

$$= \frac{\mathbf{x}\Delta}{\sigma^2} + \frac{\Delta^2}{2\sigma^2} \tag{46}$$

$$\leq \frac{3\Delta^2}{2\sigma^2} \tag{47}$$

Thus, we have for $-\Delta \leq \mathbf{x} \leq \Delta$:

$$\left| \log \left(\frac{\Pr \left(\mathbf{x} \in [\mathbf{x} - \Delta, \mathbf{x} + \Delta] \right)}{p_{\mathbf{x}}(\mathbf{x}) \cdot 2\Delta} \right) \right| \leq \frac{3\Delta^2}{2\sigma^2} \tag{48}$$

Given σ, if Δ is sufficiently small, then the approximation at $\mathbf{x}$ ($|\mathbf{x}| \leq \Delta$) is tight. □

Appendix B Probability approximation based on distribution (24)

We want to show that given marginal density function (24) the probability $\Pr(\mathbf{XW} \in [\hat{\mathbf{\Pi}}_{\mathbf{W}} - \Delta\mathbf{1}, \hat{\mathbf{\Pi}}_{\mathbf{W}} + \Delta\mathbf{1}])$ can be approximated well by using the form $p_{\mathbf{\Pi}_{\mathbf{W}}}(\mathbf{XW}) \cdot 2\Delta$ for sufficiently small Δ. As random variable $\mathbf{XW}$ in distribution (24) consists of elements that are all independent to each other, it is sufficient to show the approximation quality for a one dimensional t-distribution with degree of freedom ν:

Proposition 2 *The one dimensional r.v.* $\mathbf{x}$ *follows a t-distribution with density function*

$$p_{\mathbf{x}}(\mathbf{x}) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}.$$

Hence, the approximation of probability $Pr(\mathbf{x} \in [\mathbf{x} - \Delta, \mathbf{x} + \Delta])$ *by* $p_{\mathbf{x}}(\mathbf{x}) \cdot 2\Delta$, *has a bounded log approximation ratio:*

$$\left|\log\left(\frac{Pr(\mathbf{x} \in [\mathbf{x} - \Delta, \mathbf{x} + \Delta])}{p_{\mathbf{x}}(\mathbf{x}) \cdot 2\Delta}\right)\right| \leq \begin{cases} \frac{\Delta(2|\mathbf{x}| + \Delta)}{\mathbf{x}^2 + \nu} & : |\mathbf{x}| \geq \Delta, \\ \frac{\Delta^2}{\nu} \max\left(\frac{\nu+1}{2}, 4\right) & : |\mathbf{x}| \leq \Delta. \end{cases}$$

Thus, for given σ, ν ($\nu > 0$), and $\mathbf{x}$, if Δ is sufficiently small and $\mathbf{x}\Delta$ tends to 0, the upper bound of the log approximation ratio tends to zero. Namely, the approximation is tight.

Proof Let us first consider the case where $\mathbf{x} - \Delta > 0$. Because of the symmetry of the t-distribution, the result also applies for the case where $\mathbf{x} + \Delta < 0$. Let $1/\mathbf{Z}_\nu = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}$, we have:

- Estimation of the probability: $2\Delta \cdot \frac{1}{\mathbf{Z}_\nu}(1 + \frac{\mathbf{x}^2}{\nu})^{-\frac{\nu+1}{2}}$.
- Upper bound on the probability: $2\Delta \cdot \frac{1}{\mathbf{Z}_\nu}(1 + \frac{(\mathbf{x}-\Delta)^2}{\nu})^{-\frac{\nu+1}{2}}$.
- Lower bound on the probability: $2\Delta \cdot \frac{1}{\mathbf{Z}_\nu}(1 + \frac{(\mathbf{x}+\Delta)^2}{\nu})^{-\frac{\nu+1}{2}}$.

Then the log approximation ratio can be computed as:

1. for upper bound we have

$$\begin{aligned}\left|\log\left(\frac{2\Delta \cdot \frac{1}{\mathbf{Z}_\nu}(1 + \frac{(\mathbf{x}-\Delta)^2}{\nu})^{-\frac{\nu+1}{2}}}{2\Delta \cdot \frac{1}{\mathbf{Z}_\nu}(1 + \frac{\mathbf{x}^2}{\nu})^{-\frac{\nu+1}{2}}}\right)\right| &= \left|\log\left(\frac{\nu + \mathbf{x}^2 - 2\mathbf{x}\Delta + \Delta^2}{\mathbf{x}^2 + \nu}\right)\right| \\ &= \left|\log\left(1 + \frac{\Delta^2 - 2\mathbf{x}\Delta}{\mathbf{x}^2 + \nu}\right)\right| \\ &\leq \frac{\Delta(\Delta - 2\mathbf{x})}{\mathbf{x}^2 + \nu}\end{aligned}$$

2. and for lower bound we have

$$\left|\log\left(\frac{2\Delta\cdot\frac{1}{\mathbf{Z}_\nu}(1+\frac{(\mathbf{x}+\Delta)^2}{\nu})^{-\frac{\nu+1}{2}}}{2\Delta\cdot\frac{1}{\mathbf{Z}_\nu}(1+\frac{\mathbf{x}^2}{\nu})^{-\frac{\nu+1}{2}}}\right)\right| = \left|\log\left(\frac{\nu+\mathbf{x}^2+2\mathbf{x}\Delta+\Delta^2}{\mathbf{x}^2+\nu}\right)\right|$$
$$= \left|\log\left(1+\frac{\Delta^2+2\mathbf{x}\Delta}{\mathbf{x}^2+\nu}\right)\right|$$
$$\leq \frac{\Delta(\Delta+2\mathbf{x})}{\mathbf{x}^2+\nu}$$

By the assumption $\mathbf{x} > \Delta$, we have $\frac{\Delta(\Delta-2\mathbf{x})}{\mathbf{x}^2+\nu} \leq \frac{\Delta(\Delta+2\mathbf{x})}{\mathbf{x}^2+\nu}$, that is

$$\left|\log\left(\frac{\Pr(\mathbf{x}\in[\mathbf{x}-\Delta,\mathbf{x}+\Delta])}{p_{\mathbf{x}}(\mathbf{x})\cdot 2\Delta}\right)\right| \leq \frac{\Delta(\Delta+2\mathbf{x})}{\mathbf{x}^2+\nu}. \tag{49}$$

For given σ and $\mathbf{x}$, if Δ is sufficiently small such that $\mathbf{x}\Delta$ close to 0, then the bound $\frac{\Delta(\Delta+2\mathbf{x})}{\mathbf{x}^2+\nu}$ is close to zero. Namely, the approximation at $\mathbf{x}$ ($|\mathbf{x}| \geq \Delta$) is tight.

Remark 7 In general, for $|\mathbf{x}| \geq \Delta$, the right hand side in inequality (49) can be replaced by $\frac{\Delta(2|\mathbf{x}|+\Delta)}{\mathbf{x}^2+\nu}$

Let us now consider the case where $-\Delta < \mathbf{x} < \Delta$. Without losing generality, we assume $p(\mathbf{x}-\Delta) > p(\mathbf{x}+\Delta)$. This leads to:

- Estimation of the probability: $2\Delta\cdot\frac{1}{\mathbf{Z}_\nu}(1+\frac{\mathbf{x}^2}{\nu})^{-\frac{\nu+1}{2}}$.
- Upper bound on the probability: $2\Delta\cdot\frac{1}{\mathbf{Z}_\nu}$
- Lower bound on the probability: $2\Delta\cdot\frac{1}{\mathbf{Z}_\nu}(1+\frac{(\mathbf{x}+\Delta)^2}{\nu})^{-\frac{\nu+1}{2}}$.

Then the log approximation ratio can be computed as:

1. for upper bound we have

$$\left|\log\left(\frac{2\Delta\cdot\frac{1}{\mathbf{Z}_\nu}}{2\Delta\cdot\frac{1}{\mathbf{Z}_\nu}(1+\frac{\mathbf{x}^2}{\nu})^{-\frac{\nu+1}{2}}}\right)\right| = \left|\log\left((1+\frac{\mathbf{x}^2}{\nu})^{\frac{\nu+1}{2}}\right)\right|$$
$$\leq \frac{\nu+1}{2}\log(1+\frac{\Delta^2}{\nu})$$
$$\leq \frac{\nu+1}{2}\cdot\frac{\Delta^2}{\nu}$$

2. and for lower bound we have

$$\left|\log\left(\frac{2\Delta\cdot\frac{1}{\mathbf{Z}_\nu}(1+\frac{(\mathbf{x}+\Delta)^2}{\nu})^{-\frac{\nu+1}{2}}}{2\Delta\cdot\frac{1}{\mathbf{Z}_\nu}(1+\frac{\mathbf{x}^2}{\nu})^{-\frac{\nu+1}{2}}}\right)\right| = \left|\log\left(\frac{\nu+\mathbf{x}^2+2\mathbf{x}\Delta+\Delta^2}{\mathbf{x}^2+\nu}\right)\right|$$
$$= \left|\log\left(1+\frac{\Delta^2+2\mathbf{x}\Delta}{\mathbf{x}^2+\nu}\right)\right|$$

$$\leq \left| \log \left(\frac{\nu + 4\Delta^2}{\nu} \right) \right|$$
$$\leq \frac{4\Delta^2}{\nu}$$

For given σ and ν, if Δ is sufficiently small, then the bound $\frac{\Delta^2}{\nu} \max\left(\frac{\nu+1}{2}, 4\right)$ is close to zero. Namely, the approximation at $\mathbf{x}$ ($|\mathbf{x}| \leq \Delta$) is tight. □

References

Belkin M, Niyogi P (2003) Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput 15(6):1373–1396

Bishop CM (2006) Pattern Recogn Mach Learn. Springer, Berlin

Boley M, Mampaey M, Kang B, Tokmakov P, Wrobel S (2013) One click mining: interactive local pattern discovery through implicit preference and performance learning. In: Proceedings of the ACM SIGKDD workshop on interactive data exploration and analytics, ACM, New York, NY, USA, pp 27–35

Boumal N, Mishra B, Absil PA, Sepulchre R (2014) Manopt, a matlab toolbox for optimization on manifolds. J Mach Learn Res 15(1):1455–1459. http://www.manopt.org

Brown ET, Liu J, Brodley CE, Chang R (2012) Dis-function: learning distance functions interactively. In: IEEE VAST, IEEE, Seattle, WA, USA, pp 83–92

Cunningham JP, Ghahramani Z (2015) Linear dimensionality reduction: survey, insights, and generalizations. J Mach Learn Res 16:2859–2900

De Bie T (2011) An information theoretic framework for data mining. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining, ACM, New York, NY, USA, pp 564–572

De Bie T (2013) Subjective interestingness in exploratory data mining. In: International symposium on intelligent data analysis, Springer, Berlin, Heidelberg, pp 19–31

De Bie T, Lijffijt J, Santos-Rodriguez R, Kang B (2016) Informative data projections: a framework and two examples. In: European symposium on artificial neural networks, computational intelligence and machine learning. www.i6doc.com

Fisher RA (1936) The use of multiple measurements in taxonomic problems. Ann Eugen 7(2):179–188

Friedman JH, Tukey JW (1974) A projection pursuit algorithm for exploratory data analysis. IEEE Trans Comput 100(9):881–890

Georghiades AS, Belhumeur PN, Kriegman DJ (2001) From few to many: illumination cone models for face recognition under variable lighting and pose. IEEE Trans Pattern Anal Mach Intell 23(6):643–660

Gupta AK, Nagar DK (1999) Matrix variate distributions. CRC Press, Boca Raton

Hand DJ, Mannila H, Smyth P (2001) Principles of data mining. MIT Press, Cambridge

He X, Niyogi P (2004) Locality preserving projections. In: Advances in neural information processing systems, pp 153–160

Hotelling H (1936) Relations between two sets of variates. Biometrika 28(3/4):321–377

Hyvärinen A et al (1999) Fast and robust fixed-point algorithms for independent component analysis. IEEE Trans Neural Netw 10(3):626–634

Hyvärinen A, Karhunen J, Oja E (2004) Independent component analysis. Wiley, New York

Iwata T, Houlsby N, Ghahramani Z (2013) Active learning for interactive visualization. In: Proceedings of the sixteenth international conference on artificial intelligence and statistics, proceedings of machine learning research, vol. 31, pp 342–350

Jolliffe I (2002) Principal component analysis. Wiley Online Library

Kang B, Lijffijt J, Santos-Rodríguez R, De Bie T (2016) Subjectively interesting component analysis: data projections that contrast with prior expectations. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, ACM, New York, NY, USA, pp 1615–1624

Kohonen T (1998) The self-organizing map. Neurocomputing 21(1):1–6

Kokiopoulou E, Chen J, Saad Y (2011) Trace optimization and eigenproblems in dimension reduction methods. Numer Linear Algebra Appl 18(3):565–602

Kotz S, Nadarajah S (2004) Multivariate t-distributions and their applications. Cambridge University Press, Cambridge

Kruskal JB, Wish M (1978) Multidimensional scaling. Sage, Thousand Oaks

Lee KC, Ho J, Kriegman DJ (2005) Acquiring linear subspaces for face recognition under variable lighting. IEEE Trans Pattern Anal Mach Intell 27(5):684–698

Lijffijt J, Papapetrou P, Puolamäki K (2014) A statistical significance testing approach to mining the most informative set of patterns. Data Min Knowl Discov 28(1):238–263

Nesterov Y (2013) Introductory lectures on convex optimization: a basic course. Springer, Berlin

Ng AY, Jordan MI, Weiss Y (2002) On spectral clustering: analysis and an algorithm. In: Advances in neural information processing systems, pp 849–856

Onishchik (2011) Stiefel manifold. Encyclopedia of mathematics. http://www.encyclopediaofmath.org/index.php?title=Stiefel_manifold&oldid=12028. Accessed 21st June 2017

Paurat D, Gärtner T (2013) Invis: a tool for interactive visual data analysis. In: Machine learning and knowledge discovery in databases: European conference, ECML PKDD, Springer, Berlin, Heidelberg, pp 672–676

Peason K (1901) On lines and planes of closest fit to systems of point in space. Philos Mag 2(11):559–572

Puolamaki K, Papapetrou P, Lijffijt J (2010) Visually controllable data mining methods. In: IEEE international conference on data mining workshops, IEEE, pp 409–417

Ruotsalo T, Jacucci G, Myllymäki P, Kaski S (2015) Interactive intent modeling: information discovery beyond search. Commun ACM 58(1):86–92

Tenenbaum JB, De Silva V, Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. Science 290(5500):2319–2323

Vasilescu MAO, Terzopoulos D (2002) Multilinear analysis of image ensembles: tensorfaces. In: Proceedings of the 7th european conference on computer vision, Springer, Berlin, Heidelberg, pp 447–460

Von Luxburg U (2007) A tutorial on spectral clustering. Stat Comput 17(4):395–416

Weinberger KQ, Saul LK (2009) Distance metric learning for large margin nearest neighbor classification. J Mach Learn Res 10:207–244

Weinberger KQ, Sha F, Zhu Q, Saul LK (2006) Graph laplacian regularization for large-scale semidefinite programming. In: Advances in neural information processing systems, pp 1489–1496

Zografos K (1999) On maximum entropy characterization of Pearson's type II and VII multivariate distributions. J Multivar Anal 71(1):67–75